EFFICIENT DATA INTEGRATION TECHNIQUES IN SOME MODERN APPLICATIONS

A Dissertation
Presented to
The Academic Faculty

By

Kun Liu

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
H. Milton Stewart School of Industrial and Systems Engineering

Georgia Institute of Technology

May 2018

EFFICIENT DATA INTEGRATION TECHNIQUES IN SOME MODERN APPLICATIONS

Approved by:

Dr. Yajun Mei, Co-advisor
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Huan Xu, Co-advisor
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Xiaoming Huo
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Brani Vidakovic
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Jie Chen
Department of Population Health Sciences, Medical College of Georgia
Augusta University

Date Approved: March 30, 2018

To my parents,

for their love and encouragement.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Professor Yajun Mei, for his immense support and tremendous guidance on my research, career, and life. Without his incredible enthusiasm, valuable advices, and consistently encouragement, my dissertation would not have been accomplished. Professor Mei inspired me to the interesting and fascinating research world and taught me step by step of how to conduct research. It was him who made my graduate study enjoyable and memorable.

My sincere gratitude also goes to my co-advisor, Professor Huan Xu for his passion, support, and assistance. Professor Xu shared his deep insights through enlightening discussions, which encouraged and inspired me a lot. I am very fortunate to have Professor Xu to be my co-advisor for my Doctoral study.

I would also like to thank Professor Brani Vidakovic, Professor Xiaoming Huo, and Professor Jie Chen for serving on my thesis committee for their generous help, valuable comments, and insightful suggestions.

I am very thankful for my lab mates, friends, colleagues, and alumni at Georgia Institute of Technology. They include but are not limited to Dr. Yuan Wang, Ruizhi Zhang, Tony Yaacoub, Chen Feng, Wanrong Zhang, Yujie Zhao, Adrian Rivera Cardoso, Dr. Fang Cao, Dr. Dianpeng Wang, Dr. Minkyoung Kang, Changong Li, Xi He, Dr. Cheng Huang, Yu Cao, Chuanping Yu, and Zhehui Chen, who made my graduate life wonderful and enjoyable. I would also like to extend a special thanks to Dr. Baoguang Han and Dr. Ying Tian for supporting my internship at Biogen, Inc. It is my honor and pleasure to work with you and spend my graduate life with you.

Finally, I would like to thank for my beloved parents, for their continuous love, warm encouragement, and tremendous support. This dissertation is dedicated to them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Data science is changing our society and economy, and complicated data from heterogeneous sources is often collected in various industries such as finance, manufacturing, security, and pharmaceutical industries. The main challenge is often how to analyze these complicated data from heterogeneous sources. One useful data analysis technique is data integration that allows one to extract invaluable information from heterogeneous sources to make intelligent decisions at the global level.

This dissertation aims to develop efficient data integration techniques in some modern real-world applications. We consider four different contexts: (i) online monitoring of large-scale data streams, (ii) consensus sequential detection over distributed networks, (iii) combining different patients' responses to assess the treatment effects of new drugs, and (iv) robust statistical inference in the presence of contaminated data.

Chapter 1 investigates the problem of online monitoring large-scale data streams where an undesired event may occur at some unknown time and affect only a few unknown data streams. Existing research is either statistical inefficient or computationally infeasible. Motivated by parallel and distributed computing, we propose to develop a new information fusion technique we called the "SUM-Shrinkage" approach that is efficient and scalable. The main idea is to parallel run local detection procedures and to use the sum of the shrinkage transformation of local detection statistics as a global statistic to make a decision. The proposed shrinkage transformation approach is able to automatically filter out the unaffected data streams and only use information from affected data streams to make the decision. The usefulness of our proposed SUM-Shrinkage approach is illustrated in an example of monitoring large-scale independent normally distributed data streams when the local post-change mean shifts are unknown and can be positive or negative. Most of the material in Chapter 1 was published in a journal paper that was accepted in *Statistica Sinica* in 2018.

In Chapter 2, we consider the consensus sequential detection problem over distributed

sensor networks, in which each local sensor can only communicate local information with its immediate neighborhood sensors at each time step, and the question is how the sensors can work together to make a quick but accurate decision when testing binary hypotheses on the true raw sensor distributions. An interesting data integration technique is based on the weighted local-likelihood-ratio-statistics, which yields the Consensus-Innovation Sequential Probability Ratio Test (CISPRT) algorithm proposed by Sahu and Kar (IEEE Trans. Signal Process., 2016). Our new contribution is to present improved, non-asymptotic properties of the CISPRT algorithm for Gaussian data in term of network connectivity no matter how large the number of sensors is. Moreover, we also provide sharp upper bounds on the information loss of the CISPRT algorithm as compared to the centralized optimal SPRT algorithm in term of expected sample sizes in the asymptotic regime when Type I and II error probabilities go to 0. Numerical simulations suggest that our results are useful under the practical setting when the number of sensors is moderately large.

Chapter 3 aims to develop an efficient method that is able to combine different patients' responses to assess the treatment effects of new drugs. Our research is motivated by Biogen's ongoing Phase 3 clinical trial of a new drug "Aducanumab" for Alzheimer's disease (AD), where the primary outcome is on the *change* in the Clinical Dementia Rating-Sum of Boxes (CDR-SB) scores. The current gold standard method is the so-called responder analysis based on the two-sample proportion test, which only uses information at Month 18 and 0. This might lose detection powers because of two reasons: (i) Not every subject will have these CDR-SB scores at Month 18, due to various reasons such as missing the appointments or dropping out; (ii) it does not take advantage of the longitudinal study design when the CDR-SB scores will be collected multiple times for most subjects (e.g., at Month $0, 6, 12, 18, 24$ and 36 after the enrollment of the study). We propose to model the CDR-SB scores by the Beta distribution and to use the mixed-effects Beta regression model combining all observed CDR-SB values together to increase the detection power of the changes in the CDR-SB scores. The usefulness of our proposed models and methods is demonstrated

through the Alzheimer's Disease Neuroimaging Initiative (ADNI) database and simulation studies.

In Chapter 4 of the dissertation, we investigate the problem of robust statistical inference in the presence of contaminated data. The corrupted or contaminated data is often a big issue when we integrate data from different sources, and thus it is crucial to have a robust local inference before combining different local information together. This is still an on-going project, and here we present our preliminary research on the robust point estimations in the mixture model. Our main contribution is to consider an exponential loss function that is better to mitigate the effect of outliers and develop an asymptotic theory in a new asymptotic regime when the outlier means go to ∞ in a suitable rate as the proportion of outliers goes to 0.

CHAPTER 1

SCALABLE SUM-SHRINKAGE SCHEMES FOR DISTRIBUTED MONITORING LARGE-SCALE DATA STREAMS

1.1 Introduction

In the modern information age, one often faces the need to online monitor large-scale data streams with the aim of offering the potential for early detection of a "trigger" event. Ideally, one would like to develop a global monitoring scheme that can detect the occurring event as quickly as possible while controlling the system-wise global false alarm rate. From the statistical point of view, this is a sequential change-point detection or quickest change detection problem, which has a variety of applications such as industrial quality control, signal detection and biosurveillance. The classical version of this problem, where one monitors independent and identically distributed (iid) *univariate* or *low-dimensional multivariate* observations from a single data stream, is a well-developed area, and many classical procedures have been developed such as the Shewhart's chart ([78]), moving average control charts, Page's CUSUM procedure ([66]), Shiryaev-Roberts procedure ([79, 74]), window-limited procedures ([42]) and scan statistics ([26]). These procedures not only hold attractive theoretical properties, but also are computationally simple. See, for example, [52, 67, 68, 64, 42, 43, 41]. For a review, see the books such as [4, 69, 83].

Research has been limited in the context of monitoring large-scale data streams, especially when the occurring event might affect some, but not all, local data streams. Existing methods include the MAX-scheme (which uses the maximum of local CUSUM statistics as the global statistic, see [84]), the SUM-scheme (which uses the sum of local CUSUM statistics as the global statistic, see [59]), the mixture-schemes proposed in [97], and the simultaneous-estimation-based schemes in [92]. While the first two of these schemes are

computationally efficient but are generally statistically inefficient unless the number of affected data streams is either very small or very large, the last two schemes enjoy nice statistical properties under general settings, but are computationally infeasible for online monitoring large-scale data streams over a long time period. Our research intends to balance the tradeoff between statistical efficiency and computational efficiency when monitoring large-scale data streams.

In this chapter, we present a general and flexible approach that can provide efficient scalable global schemes when monitoring large-scale data streams. Our research is motivated by censoring sensor networks in engineering, introduced by [73] and later by [2] and [85]. Figure 1.1 illustrates the general setting of a widely used configuration of censoring sensor networks, in which the data streams $X_{k,n}$'s are observed at the remote, distributed sensors, but the final decision is made at a central location, called the fusion center. The key feature of such a network is that while taking observations at the local sensors is generally cheap and affordable, communication between remote sensors and fusion center is expensive in terms of both energy and limited bandwidth. The question then becomes how the fusion center can monitor the system effectively under the networks resource constraints in the computing power, memory and communications. An example is the National Syndromic Surveillance Program BioSense Platform at the Centers for Disease Control and Prevention (CDC), where the computing power and memory of any centralized server would be limited as compared to *daily* summary data from all state and local health departments as well as many hospitals, and thus the CDC's BioSense Platform is designed to be a distributed computing system that can detect a global level disease outbreak.

We propose to develop scalable schemes for monitoring large-scale data streams by taking advantage of parallel and distributed computing and the fact that many efficient and computationally simple local procedures are available to detect changes in local data streams. To be specific, suppose we are monitoring a large number K of data streams and, for each local data stream, an efficient local detection procedure is available based upon

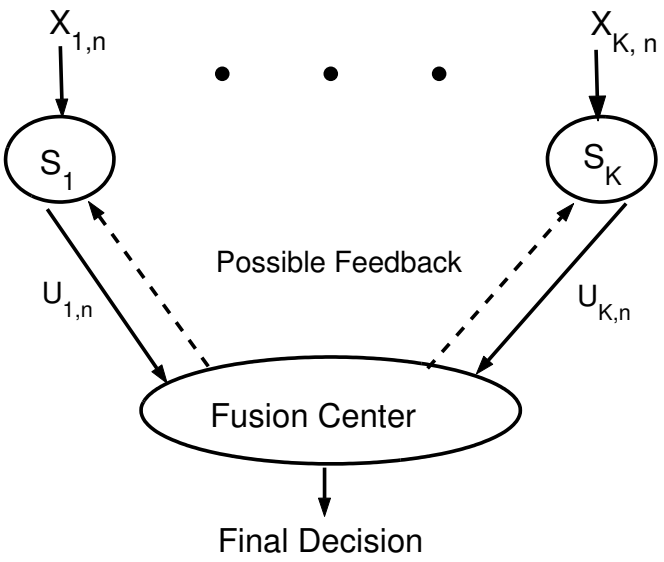

Figure 1.1: A configuration of censoring sensor networks.

some local detection statistics that can be computed recursively over time n, e.g., involving $O(1)$ computations and $O(1)$ memory requirements at each time. Our proposed methodology is to run these K local detection procedures in parallel before combining them into a global monitoring scheme. Thus the computation and memory requirements of our proposed scheme do not increase over time n, and are fixed as a function of K at each time step n when new observations are taken, thereby yielding a scalable global monitoring scheme. While the parallel local monitoring approach is interesting, a charge often made is that one loses much information at the global level by combining local detection procedures, not raw observation themselves, to make a global decision. There are two methods that combine local detection procedures together: the MAX and SUM schemes that use the maximum or sum of local CUSUM to raise a global alarm; these methods are known to be inefficient when the number of affected data streams is moderate, see [59] and [97].

In this chapter, we demonstrate that the problem is not on the parallel local monitoring approach itself, but on how to combine the local detection procedures suitably when the number of affected data streams is moderate. Our idea is to generalize the SUM scheme in [59] by introducing the shrinkage function to local detection statistics in the hope of

filtering out those unchanging local data streams. We acknowledge that there might be inherent loss of statistical efficiency in the parallel local monitoring approach as compared to the (non-recursive) global monitoring approach that uses all raw observations, e.g., see Section 1.4 for the comparison of our proposed schemes with those in [97]. The parallel local monitoring approach does allow us to develop scalable schemes, and the loss of statistical efficiency is the price we pay for the computational efficiency. A common view in the standard off-line statistical inference literature is the necessity of shrinkage for high-dimensional data in order to improve power or efficiency. Thus, from the methodology point of view, our proposed methodologies are analogous to those off-line statistical methods such as (adaptive) truncation, and soft- and hard- thresholding, see [65, 13, 14, 7], and the references there. Our motivation is different and our application to distributed quickest change detection is new.

The remainder of this chapter is organized as follows. In Section 1.2, we present some preliminaries and background information of quickest change detection or sequential change-point detection, and discuss two existing methodologies for parallel local monitoring. In Section 1.3, we propose our "SUM-shrinkage" methodology under a general setting of monitoring large-scale independent data streams, and provide general theoretical results. We exemplify our methodology in Section 1.4 for the scenario of monitoring large-scale independent normally distributed data when the post-change means of local data streams are unknown.

1.2 Preliminaries and Background

For a general setting, assume there are K independent data streams in a system.

$$\begin{aligned} &\text{Data Stream } 1: && X_{1,1}, X_{1,2}, \cdots \\ &\text{Data Stream } 2: && X_{2,1}, X_{2,2}, \cdots \\ &\cdots && \cdots \end{aligned} \tag{1.1}$$

$$\text{Data Stream } K: \qquad X_{K,1}, X_{K,2}, \cdots .$$

Initially, the system is "in control", but at some *unknown* time ν, an undesired event may occur and affect a few unknown local data streams in the sense of changing the local distributions of the $X_{k,n}$'s.

Here we assume that the online monitoring is conducted under the *unstructured* environment in the sense that we do not make any assumptions to relate the occurring event to the local data streams, see [84, 59], and [97]. Also see [45] for an application of the unstructured problem to anomaly detection in computer networks. In particular, we focus on the scenario in which the occurring event changes the local distributions of affected local data streams, and we do not aim to detect changes on the correlation between different data streams. Hence, the data $X_{k,n}$'s are assumed to be independent across different data streams, but can be flexible otherwise. For instance, the $X_{k,n}$'s may or may not be identically distributed across different local data streams, can be dependent over time within each local data stream, and can be *univariate* or *low-dimensional multivariate*. In addition, in many practical applications, the assumption of the independence across different data streams is not as restrictive as one might think, see [96] and [50], who monitor the independent residuals from some spatio-temporal models instead of dependent raw data, in applications to solar flare and hot forming processes.

For the purpose of generalization, we do not specify the kind of local changes these K data streams might have. Instead we assume that there is a local detection statistic $W_{k,n}$ (in the log-likelihood scale) for the k-th local data stream at each time step n that summarizes the evidence regarding a possible local change based on the first n local observations $(X_{k,1}, \ldots, X_{k,n})$ for each $k = 1, \ldots, K$. For instance, $W_{k,n}$ can be the well-known CUSUM or Shiryeav-Robert statistics (in the log-likelihood scale) when the local data are independent over time, or can be the recursive quasi-generalized-likelihood-ratio test in [21] when the local data are dependent from hidden Markov models. The requirements

for these $W_{k,n}$'s are that they not only should be able to detect local changes quickly, but also can be computed efficiently for our proposed scheme to be scalable. It can be highly non-trivial to construct such $W_{k,n}$'s in practice, see an example in Section 1.4.

We review the definition of a global monitoring scheme and the criteria to evaluate it under the minimax setting. A global monitoring scheme can be defined as a stopping time T with respect to the K-dimensional vector data $\{(X_{1,n}, \cdots, X_{K,n})\}_{n\geq 1}$. In particular, when $T = t$, one raises an alarm at time t to indicate that a change has occurred somewhere in the first t time steps. When monitoring K independent data streams in (1.1), even if each local false alarm rate is well controlled, the global false alarm rate can be significant when the number K of data streams is large. In the literature of sequential change-point detection, for a global monitoring scheme that raise an alarm at time T, its global false alarm rate is often evaluated by $1/\mathbf{E}^{(\infty)}(T)$, where $\mathbf{E}^{(\infty)}(T)$ is the expectation of T when the system is "in control," the Average Run Length (ARL) to false alarm. The definition of detection delay of the global monitoring scheme is more complicated. Assume that the event occurs at the unknown time ν, and the global monitoring scheme raises an alarm at time $T \geq \nu$. Then the detection delay is $T - \nu + 1$, but we must take into account of the randomness of T and the uncertainty of ν. One definition of the detection delay of T is the "worst case" delay given in [52],

$$\overline{\mathbf{E}}(T) = \sup_{\nu\geq 1} \operatorname{ess\,sup} \mathbf{E}^{(\nu)}\Big((T-\nu+1)^+ \Big| \mathcal{F}_{\nu-1}\Big). \tag{1.2}$$

Here "ess sup" is over all possible scenarios of global pre-change information $\mathcal{F}_{\nu-1} = (X_{1,[1,\nu-1]}, \ldots, X_{K,[1,\nu-1]})$, $X_{k,[1,\nu-1]} = (X_{k,1}, \ldots, X_{k,\nu-1})$ is local pre-change information up to time ν, and $\mathbf{P}^{(\nu)}$ and $\mathbf{E}^{(\nu)}$ denote the probability measure and expectation when the event occurs at time ν.

The standard minimax formulation is to find a global monitoring scheme with a stop-

ping time T that minimizes (1.2) subject to the global false alarm constraint

$$\mathbf{E}^{(\infty)}(T) \geq \gamma, \tag{1.3}$$

where $\gamma > 0$ is a pre-specified constant.

1.3 Our Proposed SUM-Shrinkage Methodology

Now we turn to our proposed methodology. Assume, for a moment, that the local detection statistics $W_{k,n}$'s (in the log-likelihood scale) have been constructed for each local k-th data streams at time n. We suggest using the global monitoring statistic of the general "SUM-shrinkage" form

$$G_n = \sum_{k=1}^{K} h_k(W_{k,n}), \tag{1.4}$$

where $h_k(\cdot) \geq 0$ are some suitable shrinkage transformation functions. Our proposed SUM-shrinkage scheme raises a global alarm at the time

$$N_G(a) = \inf\{n \geq 1 : G_n \geq a\}. \tag{1.5}$$

Our proposed $N_G(a)$ in (1.5) has two key components in its global monitoring statistic G_n in (1.4): the local detection statistic $W_{k,n}$'s; the shrinkage transformations $h_k(\cdot)$'s. Intuitively, the local detection statistics $W_{k,n}$'s should be easily computed and able to detect local changes quickly. The shrinkage functions h_k's in (1.4) play the role of dimension reduction by automatically filtering out non-changing local data streams and focusing on those local data streams that appear to be affected by the occurring event.

Our proposed "SUM-shrinkage" methodology in (1.4)-(1.5) has a broad range of applications. For instance, [60] applied the idea to develop an efficient communication policy between sensors and fusion center in the context of censoring sensor networks. Depend-

ing on which kind of local models or local changes are of interest, local detection statistic $W_{k,n}$ can be defined for such dependent observations as those from the recursive schemes in [21] for hidden Markov models, or those from the non-parametric rank-based detection schemes in [27]. Little information seems to be lost if we do not observe those local data streams with small values of the $W_{k,n}$ since they make limited contributions in the global monitoring statistic G_n in (1.4). This motivated [50] to develop an efficient adaptive sensor relocation policy when one only has ability to observe r out of K data streams at each time step. This can occur in a manufacturing process with K possible stages but only r sensors are available for monitoring. In such a problem, the order-thresholding transformation at (1.8) can be combined with missing data techniques not only to construct the global monitoring statistic G_n in (1.4) for quickest detection, but also in a greedy manner to adaptively observe those r data streams with the largest $W_{k,n}$'s values at each time step. [3] essentially tackle the similar problem of missing data, but using the hard-thresholding transformation at (1.6).

Subsection 1.3.1 contains three choices of shrinkage functions h_k at (1.4), and Subsection 1.3.2 includes some general properties of $N_G(a)$ that are related to the global false alarm constraint in (1.3). Subsection 1.3.3 discusses how to choose the tuning parameters in the shrinkage functions h_k in (1.4) when the local data streams are homogeneous.

1.3.1 Shrinkage Transformation

Evidently a suitable choice of the h_k in the SUM-shrinkage monitoring statistic G_n in (1.4) depends on the assumptions and contexts of applications. As an illustration, we list three

shrinkage transformations.

- Hard-thresholding: $h(x) = x\mathbf{1}\{x \geq b\}$ for some constant b. (1.6)
- Soft-thresholding: $h(x) = \max\{x - b, 0\}$ for some constant b. (1.7)
- Order-thresholding: $h(x) = x\mathbf{1}\{x \geq w_{(r)}\}$, where $w_{(r)}$ is the r-th largest statistic of $w_1, \cdots, w_K$. (1.8)

There are many other shrinkage functions, such as $h(x) = \exp(bx)$. By semi-Bayesian arguments, the transformation $h(x) = \log(1 - p_0 + p_0 \exp(\max(0, x)))$ was proposed by [97] in a completely different context.

To better understand the shrinkage transformations in (1.6)-(1.8), we motivate them from the communication efficiency viewpoint, first presented in [60] in the context of the censoring sensor networks in Figure 1.1. To prolong the reliability and lifetime of the network system, it is natural for the local sensors to transmit only those local detection statistics $W_{k,n}$ that are large. Specifically, at time n, the message from the sensor to the fusion center is given by

$$U_{k,n} = \begin{cases} W_{k,n}, & \text{if } W_{k,n} \geq b_k \\ \text{NULL}, & \text{if } W_{k,n} < b_k \end{cases}, \tag{1.9}$$

where $b_k \geq 0$ is the local censoring parameter at the k-th sensor (or data stream). In practice, the message "NULL" could represent that the sensor is silent.

After receiving the local sensor messages $U_{k,n}$ in (1.9), the fusion center combines them suitably to make a global decision. There are many approaches to doing so. Two schemes are based on the summation of all sensor messages $U_{k,n}$, depending on how to interpret the "NULL" values. If we treat the "NULL" values as the lower limit 0, then the fusion center

raises a global alarm at time

$$\begin{aligned} N_{hard}(a) &= \inf\Big\{n \geq 1 : \sum_{k=1}^{K} U_{k,n} \geq a\Big\} \\ &= \inf\Big\{n \geq 1 : \sum_{k=1}^{K} W_{k,n}\mathbf{1}\{W_{k,n} \geq b_k\} \geq a\Big\}. \end{aligned} \tag{1.10}$$

This scheme is referred as the hard-thresholding, since it is a special case of the global statistic in (1.4) when the shrinkage functions h_k are the hard-thresholding transformation in (1.6).

If we treat the "NULL" values as the upper limit b_k, then the fusion center computes the global monitoring statistic

$$G_n = \sum_{k=1}^{K} U_{k,n} = \sum_{k=1}^{K} \max\{W_{k,n}, b_k\} = \sum_{k=1}^{K} \max\{W_{k,n} - b_k, 0\} + \sum_{k=1}^{K} b_k,$$

which is closely related to the soft-thresholding transformation in (1.7). We can call this a soft-thresholding scheme when it raises an alarm at time

$$N_{soft}(a) = \inf\Big\{n \geq 1 : \sum_{k=1}^{K} \max\{W_{k,n} - b_k, 0\} \geq a\Big\}. \tag{1.11}$$

Here we keep the threshold of $N_{soft}(a)$ as a instead of $a - \sum_{k=1}^{K} b_k$, so that $N_{soft}(a)$ is the special case of our proposed SUM-shrinkage scheme $N_G(a)$ in (1.5) with the soft-thresholding transformation in (1.7).

A third approach occurs when the fusion center has prior knowledge that (at most) r out of K data streams will be affected by the occurring event. Such prior knowledge may be defined by the network fault-tolerant design to avoid risking failure. In this case, it is reasonable for the fusion center to order all sensor messages $U_{k,n}$'s as $U_{(1),n} \geq \ldots \geq U_{(K),n}$, and raise an alarm if the sum of the r largest $U_{k,n}$'s is too large. This is a combination of the hard-thresholding transformation in (1.6) and the order-thresholding transformation in

(1.8), and it yields a global scheme for which the stopping time is

$$N_{comb,r}(a) = \inf\Big\{n \geq 1 : \sum_{k=1}^{r} U_{(k),n} \geq a\Big\}. \tag{1.12}$$

A special case of $N_{comb,r}(a)$ in (1.12) has the order-thresholding transformation in (1.8) applied directly to the local detection statistics $W_{k,n}$ themselves. Specifically, we order the K local CUSUM statistics $W_{1,n}, \ldots, W_{K,n}$ as $W_{(1),n} \geq W_{(2),n} \geq \ldots \geq W_{(K),n}$. Then the order-thresholding scheme is defined by the stopping time

$$N_{order,r}(a) = \inf\Big\{n \geq 1 : \sum_{k=1}^{r} W_{(k),n} \geq a\Big\}, \tag{1.13}$$

which corresponds to the order-thresholding transformation in (1.8).

Based on our experience, the soft-thresholding transformation, as a continuous function, often yields smaller detection delays than the hard-thresholding transformation, a discontinuous function, in finite-sample Monte Carlo simulations. The soft- and order-thresholding transformations have comparable finite-sample performances, but the soft-thresholding transformation is computationally and theoretically simpler. We use the soft-thresholding transformation in (1.7) as a concrete demonstration, when needed.

For the soft-thresholding scheme, $N_{soft}(a)$ in (1.11), statistical intuition is a little more complicated; we provide a semi-Bayesian interpretation of why it works. At a given time n, let Z_k be the indicator of whether the distribution of the k-th local data stream changes for $k = 1, \ldots, K$. Suppose each local data stream has a prior probability π_k of being affected by the event, and that $Z_1, \ldots, Z_K$ are iid with probability mass function $\mathbf{P}(Z_k = 1) = \pi_k = 1 - \mathbf{P}(Z_k = 0)$. Treat Z_k's as the hidden states, $W_{k,n}$ representing the evidence of possible change (in logarithm scale) and applicable only when $Z_k = 1$. Then when testing $H_0 : Z_1 = \ldots = Z_K = 0$ (no change), the Log-Likelihood Ratio (LLR) statistic of the

hidden state Z_k and the observed data $X_{k,n}$ is

$$\begin{aligned} LLR(n) &= \sum_{k=1}^{K} \{Z_k(\log \pi_k + W_{k,n}) + (1 - Z_k)\log(1 - \pi_k)\} - \sum_{k=1}^{K} \log(1 - \pi_k) \\ &= \sum_{k=1}^{K} Z_k\{W_{k,n} - \log((1 - \pi_k)/\pi_k)\} \end{aligned}$$

Since the Z_k's are unobservable, it is natural to maximize $LLR(n)$ over $Z_1, \ldots, Z_K \in \{0, 1\}$. Hence, the maximum likelihood estimator of the Z_k is

$$\hat{Z}_k = \begin{cases} 1, & \text{if } W_{k,n} \geq \log((1 - \pi_k)/\pi_k) \\ 0, & \text{otherwise} \end{cases}, \quad \text{for } k = 1, \ldots, K,$$

and the generalized log-likelihood ratio is

$$\max_{Z_k's} LLR(n) = \sum_{k=1}^{K} \max\{W_{k,n} - \log((1 - \pi_k)/\pi_k), 0\},$$

exactly the form of the soft-thresholding scheme $N_{soft}(a)$ in (1.11), with $b_k = \log((1 - \pi_k)/\pi_k)$.

1.3.2 Choice of Threshold a to Satisfy The False Alarm Constraint

Given the choices of the local detection statistics $W_{k,n}$ and the shrinkage transformation $h_k(\cdot)$, an important question is how to determine the global threshold a in (1.5) so that the proposed SUM-shrinkage scheme satisfies the global false alarm constraint on γ in (1.3). This requires one to accurately characterize the relationship between the threshold a and the ARL to the false alarm $\mathbf{E}^{(\infty)}(N_G(a))$.

As the global monitoring statistic G_n in (1.4) is the sum of K (independent) random variables, one would expect that the Central Limited Theorem (CLT) would be useful when the shrinkage transformation keeps most non-zero values, e.g., the hard-thresholding or

soft-thresholding transformations in (1.6) or (1.7) when the censoring parameters b's are not large, whereas the compound Poisson process would be needed when the shrinkage transformation keeps only few non-zero values, e.g., the order-thresholding transformation in (1.8) with a not so large r value. Rigorous proofs are beyond our scope and will be investigated elsewhere.

Below we will use Chebyshev's inequalities to provide a crude relationship between the threshold a and the ARL to the false alarm $\mathbf{E}^{(\infty)}(N_G(a))$. Assume that under the pre-change hypothesis $\mathbf{P}^{(\infty)}$, for each k, the shrinkage transformation of local detection statistics, $h_k(W_{k,n})$, converge to their limit H_k^* as $n \to \infty$. We further assume that, for each $k = 1, \cdots, K$, the limit H_k^* is stochastically larger than any finite-time version $h_k(W_{k,n})$, and has a well-defined log-moment generating function

$$\psi_k(\theta) = \log \mathbf{E}^{(\infty)} \exp(\theta H_k^*) \tag{1.14}$$

for some $\theta \geq 0$.

Theorem 1.1. *Assume the $\psi_k(\theta)$ are well-defined for all $\theta \in \Theta$, a sub-interval of $[0, \infty)$, for all $k = 1, \cdots, K$. Then,*

$$\mathbf{E}^{(\infty)}(N_G(a)) \geq \frac{1}{4} \exp\Big(\theta a - \sum_{k=1}^{K} \psi_k(\theta)\Big) \tag{1.15}$$

for all $\theta \in \Theta$, and a choice of threshold

$$a = \inf_{\theta \in \Theta} \left(\frac{1}{\theta} \big(\log(4\gamma) + \sum_{k=1}^{K} \psi_k(\theta) \big) \right) \tag{1.16}$$

guarantees that $N_G(a)$ in (1.5) satisfies the global false alarm constraint γ in (1.3).

Proof: Relation (1.16) follows directly from (1.15), and it suffices to show that (1.15) holds for any $\theta \in \Theta$. By the definition of $N_G(a)$ in (1.5) and the use of Chebyshev's

inequality twice, once to $N_G(a) \geq 0$ and the second to $\sum_{k=1}^{K} H_k^*$, for any $x > 0$

$$\begin{aligned}
\mathbf{E}^{(\infty)}(N_G(a)) &\geq x\mathbf{P}^{(\infty)}\Big(N_G(a) \geq x\Big) \\
&= x\Big[1 - \mathbf{P}^{(\infty)}\Big(N_G(a) < x\Big)\Big] \\
&= x\Big[1 - \mathbf{P}^{(\infty)}\Big(\sum_{k=1}^{K} h_k(W_{k,n}) \geq a \text{ for some } 1 \leq n \leq x\Big)\Big] \\
&\geq x\Big[1 - x\mathbf{P}^{(\infty)}\Big(\sum_{k=1}^{K} H_k^* \geq a\Big)\Big] \\
&\geq x\Big[1 - xe^{-\theta a}\mathbf{E}^{(\infty)} \exp\Big(\theta \sum_{k=1}^{K} H_k^*\Big)\Big] \\
&= x\Big[1 - xe^{-\theta a} \exp\Big(\sum_{k=1}^{K} \psi_k(\theta)\Big)\Big].
\end{aligned}$$

Here the second inequality follows from the assumption that H_k^* is stochastically larger than $h_k(W_{k,n})$, and the last equation uses the assumption that these K data streams are independent across different data streams. For any $u > 0$, the function $x(1 - xu)$ is maximized at $x = 1/(2u)$ with the maximum value $1/(4u)$. This completes the proof of (1.15). □

The results in Theorem 1.1 are non-asymptotic, and hold for any K and γ. To demonstrate their usefulness, consider a concrete homogeneous case when the $W_{k,n}$s are identically distributed over k under the pre-change hypothesis, and all local data streams use the same soft-thresholding transformation (1.7). We suppress the script k and derive the log-moment generating function $\psi(\theta)$ in (1.14) for the soft-thresholding transformation $h(W_n) = \max(W_n - b, 0)$ for large b. We further assume that, as $n \to \infty$, the local detection statistic W_n converges to an asymptotically exponentially distributed variable W^* under the pre-change hypothesis,

$$\mathbf{P}^{(\infty)}(W^* > x) \approx \lambda e^{-x}, \tag{1.17}$$

for some constant $\lambda > 0$. A non-asymptotic result is often true for many local detection

statistic W_n such as CUSUM: for *any* $x > 0$,

$$\mathbf{P}^{(\infty)}(W^* > x) \leq e^{-x}, \tag{1.18}$$

see Appendix 2 of [80]. Under (1.17), we have $\mathbf{P}^{(\infty)}(W^* \leq b) = 1 - \lambda e^{-b}$ for large b. Combining the definition of $\psi(\theta)$ in (1.14) with the fact that $H^* = 0$ whenever $W^* \leq b$ yields that

$$\begin{aligned} \psi(\theta) &= \log \mathbf{E}^{(\infty)} \exp(\theta H^*) = \log[\mathbf{P}^{(\infty)}(W^* \leq b) + \int_b^\infty e^{\theta(x-b)} \lambda e^{-x} dx] \\ &= \log\left(1 + \frac{\theta\lambda e^{-b}}{1-\theta}\right). \end{aligned} \tag{1.19}$$

Clearly, $\psi(\theta)$ is well-defined over $\theta \in \Theta = [0, 1)$. If we further assume that b is large, or equivalently, λe^{-b} is small, using the approximation $\log(1 + x) \approx x$ yields that $\psi(\theta) \approx \theta\lambda e^{-b}/(1 - \theta)$. Thus the term inside the infimum in (1.16) is

$$\frac{1}{\theta}\left(\log(4\gamma) + K\psi(\theta)\right) \approx \frac{1}{\theta}\log(4\gamma) + \frac{1}{1-\theta}(K\lambda e^{-b}).$$

As $A/\theta + B/(1-\theta)$ has a minimum value $(\sqrt{A} + \sqrt{B})^2$ over $0 \leq \theta \leq 1$ for any $A, B > 0$, (1.16) in Theorem 1.1 gives

$$a \approx \left(\sqrt{\log(4\gamma)} + \sqrt{K\lambda e^{-b}}\right)^2. \tag{1.20}$$

In(1.20) we see the challenges of monitoring large-scale data streams: the asymptotic expression of a in (1.20) depends on the asymptotic relationship between $\log(\gamma)$ and $K\lambda e^{-b}$. When $\log(\gamma) >> K$, we have the classical result on the threshold of $a = (1 + o(1))\log(\gamma)$, see [52]. When $K\lambda e^{-b} >> \log(\gamma)$, we have

$$a \approx K\lambda e^{-b} + 2\sqrt{K\lambda e^{-b}}\sqrt{\log\gamma}. \tag{1.21}$$

This suggests that $K\lambda e^{-b}$ plays a dominant role to determine the threshold a for $N_G(a)$ to satisfy the false alarm constraint γ in (1.3) when b is large and $Ke^{-b} >> \log(\gamma)$.

1.3.3 The Choice of Censoring Parameters

In this subsection, we discuss the optimal choice of the censoring parameters b_k in (1.9). For illustration and simplicity we consider the homogeneous case, $b_k \equiv b$, when local data streams are identically distributed for different k, e.g., relations (1.17), (1.18), and $\psi_k(\theta) \equiv \psi(\theta)$ in (1.19) hold for all $k = 1, \cdots, K$. We provide two optimal choices of the censoring parameter b for the soft-thresholding scheme $N_{soft}(a)$ in (1.11): one from the communication efficiency aspect, and the other from the statistical efficiency aspect. It turns out that they are closely related.

Assume that the average fraction of transmitting sensors at any time step is restricted to be at most $\eta \in (0, 1)$ when no change occurs. In this case, when no event occurs, the average fraction of transmitting sensors at any time step n is

$$\begin{aligned} \frac{1}{K}\sum_{k=1}^{K} \mathbf{P}^{(\infty)}(U_{k,n} \neq \text{NULL}) &= \frac{1}{K}\sum_{k=1}^{K} \mathbf{P}^{(\infty)}(W_{k,n} \geq b) \\ &\leq \frac{1}{K}\sum_{k=1}^{K} \exp(-b) \leq \exp(-b), \end{aligned}$$

where the second-to-last inequality follows from (1.18). Thus a choice of

$$b_{opt,1} = \log(\eta^{-1}) \tag{1.22}$$

will guarantee that on average, at most $100\eta\%$ of K sensors transmit messages at any given time when no event occurs. When η is small, one can use the refined asymptotic approximation (1.17), instead of the non-asymptotic bound (1.18), in the above arguments. Then the $b_{opt,1}$ can further improved as $b^*_{opt,1} = \log(\lambda/\eta)$ under the communication constraint.

Next, we choose the censoring parameter b based on the statistical efficiency considerations in the scenario when w_0 out of K local data streams are affected. Intuitively, when the global threshold value a is *given*, our proposed scheme $N_{soft}(a)$ in (1.11) is increasing as a function of the censoring parameter $b_k \equiv b$, and a larger value of b implies both larger ARL to false alarm and larger detection delays. Hence, subject to the false alarm constraint γ in (1.3), different global threshold values a are needed for these schemes with different b, and thus it is natural to find the censoring parameter b that yields the smallest detection delay $\overline{\mathbf{E}}(T)$ in (1.2).

We assume that those affected local streams have the same post-change statistical properties in the sense that the detection delay of a local scheme $N_k(c) = \inf\{n \geq 1 : W_{k,n} \geq c\}$ is $(1 + o(1))c/I$ for some constant $I > 0$ as $c \to \infty$. This assumption is general and holds for many local detection statistics including CUSUM, see [52]. Then the detection delay of the soft-thresholding scheme $N_{soft}(a)$ in (1.11) is bounded above by

$$(1 + o(1))\frac{1}{I}\left(b + \frac{a}{w_0}\right). \tag{1.23}$$

To see this, at time step n, if $w_{k,n} \geq b + a/w_0$ for all of those w_0 affected local data streams, then $N_{soft}(a) \leq n$ since $\sum_{k=1}^{K} \max(w_{k,n} - b, 0) \geq w_0(a/w_0) = a$. Relation (1.23) follows at once from the detection delays of $N_k(c)$ with $c = b + a/w_0$ for those w_0 affected data streams, and similar ideas have been applied in the proof of Theorem 3 in [57] when the $W_{k,n}$ are local CUSUM statistics.

If we keep only on the first-order major term of a in (1.21), plugging it into (1.23) yields that the detection delay of the soft-thresholding scheme $N_{soft}(a)$ in (1.11) (up to the first-order) is

$$\frac{1}{I}\left(b + \frac{K\lambda e^{-b}}{w_0}\right).$$

Taking derivatives with respect to b, and setting it to 0, the detection delay bound is mini-

mized when $K\lambda e^{-b} = w_0$, so the optimal b value (up to first-order) is given by

$$b_{opt,2} = \log \frac{\lambda K}{w_0}, \tag{1.24}$$

where $\lambda > 0$ is the constant in (1.17) that only depends on the asymptotic properties of the $W_{k,n}$.

When we have prior knowledge that w_0 local data streams are affected but we do not know which ones, it is reasonable to assume that each local data stream has the same probability $\pi = w_0/K$ of being affected. By the semi-Bayesian interpretation of the soft-thresholding transformation in Subsection 1.3.1, the local censoring parameters b_k's should be chosen as $b_k = \log((1-\pi)/\pi) = \log((K-w_0)/w_0)$, which is asymptotically equivalent to (1.24) when the fraction of affected data stream $w_0/K \to 0$.

A direct comparison of (1.22) and (1.24) suggests that the two optimal b values are asymptotically equivalent if we set $\eta = w_0/K$. Moreover, by (1.9) and (1.17), when there are no changes, the average fraction of transmitting sensors at any time step n is

$$\begin{aligned} \frac{1}{K}\sum_{k=1}^{K} \mathbf{P}^{(\infty)}(U_{k,n} \neq \text{NULL}) &= \mathbf{P}^{(\infty)}(U_{k,n} \neq \text{NULL}) = \mathbf{P}^{(\infty)}(W_{k,n} \geq b_{opt,2}) \\ &= \lambda e^{-b_{opt,2}} = w_0/K. \end{aligned} \tag{1.25}$$

This demonstrates a simple but useful equivalence relationship between communication efficiency and statistical efficiency: if we want to optimize the detection delay performance (up to first-order) when w_0 data streams are affected, then it is best to design the schemes that on average allow w_0 out of K local data streams to transmit local detection statistics to the fusion center when no change event occurs (and possibly more than w_0 data streams when a change occurs). Due to this equivalence, in our simulations, the censoring parameter b is chosen based on (1.22), which is non-asymptotic and easier to compute.

1.4 An Example: Unknown Post-Change Normal Means

Suppose that we are monitoring K independent normally distributed data streams $X_{k,n}$ in (1.1). Initially, the data $X_{k,n}$ are iid $N(0,1)$. At some unknown time ν, the distribution of the k-th local data stream might change to $N(\mu_k, 1)$ if affected. We do not know which subset of local data streams are affected, and here another new challenge is that we do not know the values of the post-change means μ_k's when affected. We want to develop a system-wise online monitoring scheme that can detect the change as soon as possible, subject to the global false alarm constraint γ in (1.3).

[97] investigated this problem under the assumption that the post-change mean $\mu_k > 0$ for all k. By assuming that the fraction p_0 of affected data stream is known, the scheme they proposed was motivated from a semi-Bayesian approach; it is defined by

$$T_{XS}(a, p_0) = \inf \left\{ n \geq 1 : \max_{0 \leq i < n} \sum_{k=1}^{K} \log(1 - p_0 + p_0 \exp\left[\left(U^+_{k,n,i}\right)^2 / 2 \right] \geq a \right\}, \quad (1.26)$$

where $U^+_{k,n,i} = \max\left(0, \sum_{j=i+1}^{n} X_{k,j}\right)/\sqrt{n-i}$ for all $0 \leq i < n$ and $1 \leq k \leq K$. One can also simplify the memory requirement by keeping a large window of the most recent observations. [92] proposed a global Shiryaev-Robert procedure by simultaneously estimating all K unknown post-change means μ_k via shrinkage estimation. These schemes are not scalable, and not suitable in the context of censoring sensor networks in Figure 1.1. The implementation of their schemes requires the fusion center to have full access to all data streams at each time step.

It has been an open problem to develop a scalable global monitoring scheme that is able to detect both positive and negative local mean shifts for affected local data streams. Part of the reason is that for the K local data streams, there are 2^K potential different combinations of positive or negative local shifts, and not feasible for large K. Here we illustrate how to tackle this problem based on our proposed SUM-shrinkage statistics in (1.4). We need a suitable local detection statistic $W_{k,n}$ that can be easily computed and has the ability to

detect both positive and negative local mean shifts. if the local detection statistics $W_{k,n}$s are defined, we can use any shrinkage transformation to develop a global monitoring scheme.

In this section, we consider the soft-thresholding scheme $N_{soft}(a)$ in (1.11). For simplicity, we assume that all censoring parameters b_k in (1.11) are the same, $b_k \equiv b_1$ for some constant $b_1 > 0$. Our focus is how to construct the local detection statistics $W_{k,n}$'s suitably.

Subsection 1.4.1 provides an overview of our proposed soft-thresholding scheme in (1.11) that only uses a fixed number of $6K$ registers to store all past information and involves $O(K)$ computations at each given time step n. Simulation results are summarized in Subsection 1.4.2.

1.4.1 Our Proposed Local Detection Statistics $W_{k,n}$

We are interested in detecting both positive and negative local mean shifts for affected data streams, we propose to extend the detection statistic W_n of [53] from one-sided to two-sided. As detecting negative local mean shift of the $X_{k,n}$ is equivalent to detecting positive local mean shift of the $-X_{k,n}$, we propose the local detection statistic for each local data stream at time n,

$$W_{k,n} = \max\left(W_{k,n}^{(1)}, W_{k,n}^{(2)}\right). \tag{1.27}$$

Here $W_{k,n}^{(1)}$ and $W_{k,n}^{(2)}$ are the local detection statistics of [53] for detecting positive and negative mean shifts, respectively:

$$\begin{aligned} W_{k,n}^{(1)} &= \max\left(W_{k,n-1}^{(1)} + \hat{\mu}_{k,n}^{(1)} X_{k,n} - \frac{1}{2}(\hat{\mu}_{k,n}^{(1)})^2, 0\right), \\ W_{k,n}^{(2)} &= \max\left(W_{k,n-1}^{(2)} + \hat{\mu}_{k,n}^{(2)} X_{k,n} - \frac{1}{2}(\hat{\mu}_{k,n}^{(2)})^2, 0\right), \end{aligned} \tag{1.28}$$

where

$$\hat{\mu}_{k,n}^{(1)} = \max\Big(\rho, \frac{s + S_{k,n}^{(1)}}{t + T_{k,n}^{(1)}}\Big) > 0, \qquad \hat{\mu}_{k,n}^{(2)} = \min\Big(- \rho, \frac{-s + S_{k,n}^{(2)}}{t + T_{k,n}^{(2)}}\Big) < 0, \tag{1.29}$$

and for $j = 1, 2$, the sequences $(S_{k,n}^{(j)}, T_{k,n}^{(j)})$ are defined over n recursively as

$$\begin{pmatrix} S_{k,n}^{(j)} \\ T_{k,n}^{(j)} \end{pmatrix} = \begin{cases} \begin{pmatrix} S_{k,n-1}^{(j)} + X_{k,n-1} \\ T_{k,n-1}^{(j)} + 1 \end{pmatrix} & \text{if } W_{k,n-1}^{(j)} > 0 \\ \begin{pmatrix} 0 \\ 0 \end{pmatrix} & \text{if } W_{k,n-1}^{(j)} = 0 \end{cases} \tag{1.30}$$

Here the $\hat{\mu}_{k,n}^{(1)}$ and $\hat{\mu}_{k,n}^{(2)}$ in (1.29) are the estimates of the post-change mean when restricted to the positive and negative values, respectively, under the assumption that $|\mu| \geq \rho$. The two-sided local detection statistic $W_{k,n}$ in (1.27) is always nonnegative for any k at any time step n, and it is large when there is a local mean shift no matter whether such mean shift is positive or negative.

The proposed soft-thresholding scheme $N_{soft}(a)$ in (1.11) can be easily implemented in the censoring sensor network context by parallel computing the K local detection statistics $W_{k,n}$'s recursively through (1.27)-(1.30) at the local sensor levels. We can use $6K$ registers to adaptively store all past information at each time step after observing new data: $(S_k^{(j)}, T_k^{(j)}, W_k^{(j)})$ for $j = 1, 2$ and $k = 1, 2 \cdots, K$. At any given time step n, we can first update the $4K$ registers in $(S_k^{(j)}, T_k^{(j)})$ using the past data and compute the $2K$ estimates $\hat{\mu}_k^{(j)}$ of the post-change means μ_k's. Then after we observe new observations, $(X_{1,n}, \cdots, X_{K,n})$, we only need update the $2K$ registers $W_k^{(j)}$'s and compute the values of K local detection statistics W_k's, which allows us to easily compute the global monitoring statistic G. Including the $3K$ intermediate variables $(\hat{\mu}_k^{(j)}, W_k)$ and the global monitoring statistic G, the proposed scheme only needs $9K + 1$ registers to adaptively store all relevant information and involves $O(K)$ computations at any given time step n. Our scheme

can be implemented in censoring sensor networks where most computations are done at the remote sensors. Hence, our proposed scheme is scalable and can be easily implemented to online monitor large-scale data streams over a long time period.

1.4.2 Simulation Results

In this subsection, we report the numerical simulation results of the soft-thresholding scheme $N_{soft}(a)$ in (1.11) when the local detection statistics $W_{k,n}$'s are defined recursively through (1.27)-(1.30), and the censoring parameters $b_k \equiv b_1$ for all k. For the purpose of comparison, we follow Xie and Siegmund [97] to assume that there are $K = 100$ independent normal data streams. For each $k = 1, \cdots, K$, the data $X_{k,n}$'s of the k-th data stream are iid $N(0,1)$ before the change, but are iid $N(1,1)$ after the k-th data stream is affected by the occurring event.

In our simulations, we considered six schemes: the Xie and Siegmund schemes $T_{XS}(a, p_0)$ in (1.26) with $p_0 = 1$ and 0.1, and four schemes employed our proposed soft-thresholding schemes $N_{soft}(a)$ in (1.11) with censoring parameters: $b_1 = 0, 0.5, \log(10), \log(100)$. The three non-zero b_1 values imply that on average at most $\exp(-b_1) \approx 60.7\%, 10\%$ and 1% out of 100 local data streams produce significant $W_{k,n}$ values to the global monitoring statistic G_n when there are no changes. When computing the local detection statistics $W_{k,n}$ in (1.27), we set $\rho = 0.25, t = 4$ and $s = 1$ as in [53].

For each of these schemes, we first numerically searched the threshold a to satisfy the global false alarm constraint γ in (1.3). Two values of γ were considered: $\gamma = 5000$, so that we can compare with those results from [97]; $\gamma = 5 \times 10^4$ to see the effect of false alarm constraint γ on the detection delays of our schemes. We are unable to numerically find the global threshold a of the Xie and Siegmund scheme for the case $\gamma = 5 \times 10^4$ in a reasonable time, and there we only report the performance of our proposed schemes. For the detection delays, we considered various post-change hypotheses and, for each post-change hypothesis, we simulated the $\mathbf{E}(T(a))$ when the event occurs at time $\nu = 1$, and

Table 1.1: A comparison of detection delays when the change is instantaneous and the post-change mean $\mu_k = 1$ if affected. The smallest and largest standard errors of the schemes are reported under each post-change hypothesis based on 2500 repetitions in Monte Carlo simulations.

γ		# local data streams affected								
		1	3	5	8	10	20	30	50	100
	Smallest standard error	0.19	0.08	0.06	0.04	0.03	0.02	0.01	0.01	0.00
	Largest standard error	0.40	0.14	0.08	0.05	0.04	0.03	0.02	0.02	0.01
5000	Xie and Siegmund's schemes $T_{XS}(a, p_0)$ in (1.26)									
	$T_{XS}(a = 53.5, p_0 = 1)$	52.4	18.3	11.1	7.1	5.7	2.9	2.0	1.2	1.0
	$T_{XS}(a = 19.5, p_0 = 0.1)$	31.1	13.4	9.2	6.7	5.7	3.5	2.5	1.8	1.0
	Soft-thresholding Schemes $N_{soft}(a)$ in (1.11)									
	$N_{soft}(a = 127.86, b_1 = 0)$	75.0	35.4	25.2	18.5	16.0	10.3	8.1	6.1	4.1
	$N_{soft}(a = 84.91, b_1 = 0.5)$	72.1	33.9	24.1	17.7	15.3	10.0	7.9	6.0	4.2
	$N_{soft}(a = 24.01, b_1 = \log(10))$	45.8	22.0	16.4	12.8	11.5	8.5	7.3	6.1	5.0
	$N_{soft}(a = 7.88, b_1 = \log(100))$	29.0	17.2	14.2	12.0	11.2	9.2	8.3	7.3	6.4
5×10^4	Soft-thresholding Schemes $N_{soft}(a)$ in (1.11)									
	$N_{soft}(a = 136.07, b_1 = 0)$	89.0	39.9	27.9	20.2	17.4	11.1	8.7	6.5	4.4
	$N_{soft}(a = 92.79, b_1 = 0.5)$	85.7	38.2	26.8	19.4	16.7	10.7	8.4	6.3	4.4
	$N_{soft}(a = 29.05, b_1 = \log(10))$	55.1	25.3	18.4	14.1	12.6	9.1	7.8	6.5	5.2
	$N_{soft}(a = 11.11, b_1 = \log(100))$	35.5	19.7	16.0	13.4	12.4	10.0	8.9	7.9	6.8

used this as an estimate of the detection delay $\mathbf{D}(T(a))$. All simulated values were based on 2500 Monte Carlo runs.

Table 1.1 summarizes detection delays when the change is instantaneous if a local data stream is affected. For the Xie and Siegmund scheme $T_{XS}(a, p_0)$ in (1.26), our simulated detection delay results are slightly different from their reported results in their paper, possibly because our simulation was based on 2500 runs instead of the 500 runs in their paper. The Xie and Siegmund schemes $T_{XS}(a, p_0)$ in (1.26) involve expensive computations, and require the fusion center to have full access to all raw data. Thus it is not surprising that their schemes have smaller detection delays than our schemes. The Xie and Siegmund schemes are not scalable and cannot be implemented in the context of distributed monitoring in censoring sensor networks. Our proposed schemes can be easily implemented by parallel computing in a recursive manner at the local sensors level and thus are scalable.

All simulations were done on a Windows 8 Laptop with Intel i7-4700MQ CPU 2.40GHz using MATLAB R2013b. For each row of Table 1.1, the most time consuming part was to search for the global threshold a so that $\mathbf{E}^{(\infty)}(T(a)) \approx \gamma$. When $\gamma = 5000$, it took

about 8 minutes to find such a from *a range of values* for our proposed schemes based on 2500 Monte Carlo runs (the time is shorter if our initial guess range of a is closer). For the Xie and Siegmund scheme, for the *given* global threshold a around 53.5 provided in their paper, it took about one and a half hours on average to finish one Monte Carlo simulation run. If we did not know $a \approx 53.5$ and wanted to try 10 different values of a's by bisection method based on 2500 Monte Carlo runs for each a, it would have taken about $10 \times 1.5 \times 2500 = 37500$ computer hours for the case of $\gamma = 5000$. When $\gamma = 5 \times 10^4$, it took us about one hour to find the global threshold a for our proposed schemes, but we are unable to numerically implement the Xie and Siegmund schemes. Once the global threshold a is found, it is straightforward to simulate the detection delays in Table 1.1. When $\gamma = 5000$, our proposed schemes are at least 10 times faster than those of Xie and Siegmund. For instance, when exactly one data stream is affected, it took 4.94 seconds to simulate the detection delay of our proposed schemes, and 41.02 seconds to simulate theirs. The computational advantage of our proposed schemes is evident.

CHAPTER 2
IMPROVED PERFORMANCE PROPERTIES OF THE CISPRT ALGORITHM FOR DISTRIBUTED SEQUENTIAL DETECTION

2.1 Introduction

Distributed online learning becomes increasingly important in many real-world applications such as cognitive radio networks [48, 49], social recommender systems [86, 99], natural language processing [25]. Under a general setting, there are N sensors or agents taking raw observations over time in a system, and each local sensor can only communicate its local information with the immediate neighbors at each time. Such local information communication can be conducted adaptively or sequentially over time so that sensors can work together to reach consensus quickly. The advantages of distributed settings are to protect intrinsic privacy of sensitive data [99], increase computational capacity [22, 71, 100], and mitigate collection and storage burden of modern large datasets [54, 72].

There are many important distributed online learning problems in engineering and statistics, and one of them is the distributed sequential detection, see [6, 90, 101], where the distributed sensors work together to quickly and correctly decide which is the true underlying probability measure or model for raw sensor observations. Had the local sensors been able to send local information to a central location, often called the fusion center, for further analysis, extensive research has been done along two distinct directions. The first one is when the fusion center has access to all raw sensor observations, which is the centralized sequential detection problem. This is well studied in the classical subfield of sequential analysis in statistics [4, 80, 83, 91]. In particular, the optimal centralized procedure is the well-known Sequential Probability Ratio Test (SPRT), see [91]. The other direction is the decentralized sequential detection, where the local sensors send quantized sensor messages

to the fusion center to make a global decision, see [15, 28, 58, 88].

Research is rather limited in the distributed sequential detection where there is no fusion center and the local sensors need to work together to make a decision: very few efficient algorithms have been proposed partly because it involves complicated communication strategies between local sensors and their neighborhood sensors. One exception is the Consensus-Innovation SPRT (CISPRT) algorithm developed in [75] that is based on the weighted average of local log-likelihood ratio tests, see [37, 39, 38, 36, 55] for the motivation and more background. The CISPRT algorithm is novel and interesting, as each local sensor utilizes local information not only from itself and its immediate neighbor sensors, but also from remote connected sensors that are 2-hop or more hops away from itself. Also see [46] for an interesting generalization of the CISPRT algorithm under the fixed q-round message passing protocol and see [76] for the extension in composite hypothesis testing problems.

Intuitively, the performance of distributed algorithms including the CISPRT will depend on the neighborhood structure of local senors, or the network connectivity. For any pre-specified neighborhood structure of local senors, Sahu and Kar [75] characterized the various performance properties of the CISPRT for Gaussian data. In particular, for the CISPRT satisfying the error probability constraint ϵ, explicit lower and upper bounds were derived on its performance properties $h(\epsilon)$ such as the expected sample sizes and the information loss with respect to the optimal centralized SPRT: $L(\epsilon) < h(\epsilon) < U(\epsilon)$, which hold non-asymptotically for *any* network structure regardless of the number of sensors. These are remarkable non-asymptotic results on sequential detection in the high-dimensional setting, as the explicit bounds $L(\epsilon)$ and $U(\epsilon)$ clearly characterize the effects of the network structure and the dimension (or the number of sensors). Unfortunately, these bounds are too loose in the special centralized setting when each local sensor is connected to all other sensors and the CISPRT becomes the well-known centralized SPRT: the ratios $U(\epsilon)/L(\epsilon)$ converge to $5/4$ and $10/7$ for the expected sample size and information loss, respectively,

in the asymptotic setting as the error probability constraint $\epsilon \to 0$. This led us to raise an open problem whether one can derive better lower and/or upper bounds.

The main objective of this chapter is to provide a positive answer to this open problem. Our focus is to improve the upper bounds in [75] for the CISPRT algorithms, and also show that our derived upper bounds are asymptotically tight up to first order. Note that it is mathematically challenging to provide an accurate analysis on the performance properties of sequential detection procedures regardless of the number N of sensors. The standard techniques in sequential detection or sequential hypothesis testing are renewal theory and overshoot analysis, but they are designed for the fixed dimensional setting and are inappropriate in our context for any network structure regardless of the dimension N (or the number N of sensors), since the corresponding results involve implicit overshoot constants that will be exponentially increasing as a function of the dimension N.

Our main scientific contributions are two-fold. From the technique viewpoint, we develop a tail probability analysis technique that is able to derive sharp information bounds that are not only comparable to the classical techniques in the one or low dimensional setting (i.e., when there are one or very few sensors), but also much better in the high-dimensional setting (i.e., when there are a large number of sensors). Our proposed technique is to extend the finite sum of tail probabilities in [75] to the infinite sum, and provides a new and useful tool that is able to provide accurate performance analysis in the sequential detection context. From the network application viewpoint, we derive refined, non-asymptotic upper bounds on the performance properties of the CISPRT algorithm for Gaussian data under any pre-specified neighborhood structure of local senors regardless of the number N of sensors, as compared to those in [75]. In particular, our derived upper bounds are asymptotically equivalent to the lower bounds in the sense that $U(\epsilon)/L(\epsilon) \to 1$ for the special centralized setting as the error probabilities constraint $\epsilon \to 0$. Our results indicate that the more the number of sensors or the sparser the network neighborhood connectivity is, the larger the information loss is, i.e., the larger expected sample size is needed

to achieve the desired Type I and II error probabilities.

The remainder of this chapter is organized as follows. In Section 2.2, we present the formulation of the distributed sequential detection problem, the CISPRT algorithm proposed in [75], and the background materials for spectral graph theory. In Section 2.3, we present our main theoretical results on the refined performance properties of the CISPRT. Simulation studies results are presented in Section 2.4, and the detailed proof of our main theorem, Theorem 2.1, is provided in Section 2.5. Some conclusion remarks are included in Section 2.6.

2.2 Preliminaries and Background

2.2.1 Distributed Sequential Detection Problems

Consider a network system of N sensors that take observations over time. At each time step $t = 1, 2, \cdots$, the i-th sensor observes an observation $y_i(t)$ for $i = 1, \ldots, N$. There are two hypotheses on the distributions of the local sensor observations $y_i(t)$'s. Under the null hypothesis H_0, the $y_i(t)$'s are $N(-\mu, \sigma^2)$ and under the alternative hypothesis H_1, the $y_i(t)$'s are $N(\mu, \sigma^2)$. Here the sensor observations $y_i(t)$'s are assumed to be independent and identically distributed (iid) over time and across sensors, conditional on each hypothesis. Note that Guassian distributions are one of the most widely used models in signal processing and many other applications, and by the linear additive properties of Gaussian models, our problem is equivalent to a more familiar problem in the additive white Gaussian noise channel of utilizing the observations $y_i^*(t) = y_i(t) + \mu$ to test hypotheses $H_0 : N(0, \sigma^2)$ (i.e., white noises) against $H_1 : N(2\mu, \sigma^2)$ (i.e., a signal exists). Here we follow [75] to adopt the current notation so as to simplify the technical presentations and proofs.

Under the distributed sequential detection setting, the objective is for each local sensor to work with its neighborhood sensors to make a quick and accurate decision on which of these two hypotheses is true. In particular, each local sensor can only communicate

its local information with its (one-hop) neighborhood sensors. Here we assume that the neighborhood structure of sensors is pre-specified, and can be represented as an undirected graph $G = (V, E)$: the i-th vertex in V represents the i-th sensor, and there is an edge between the i-th vertex/sensor and the j-th vertex/sensor, i.e., $(i, j) \in E$, if and only if the corresponding sensors are neighbors and can communicate local information with each other. Here we assume that the graph $G = (V, E)$ is simple, i.e., without self loops and multiple edges. For the i-th sensor, its neighborhood is given by $\Omega_i = \{j \in V | (i, j) \in E\}$, and its degree is given by the cardinality $d_i = |\Omega_i|$. See Mesbahi and Egerstedt [61] for more graph theoretic methods in network systems.

For a distributed sequential procedure D, it consists of $(T_i, \delta_i)_{i=1}^N$, where T_i is the number of time steps the i-th local sensor needed to make a local decision $\delta_i \in \{0, 1\}$. Here T_i is a local stopping time in the sense that $\{T_i = t\}$ depends on the information from the i-th local sensor as well as its neighborhood up to time t. The local decision $\delta_i = 0$ or 1 means that the i-th local sensor accepts H_0 or H_1, respectively.

The performance of a distributed sequential procedure $D = (T_i, \delta_i)_{i=1}^N$ is evaluated by its local expected sample sizes, $\mathbf{E}_1[T_i]$ and $\mathbf{E}_0[T_i]$, and its local error probabilities, $\mathbf{P}_0[\delta_i = 1]$ and $\mathbf{P}_1[\delta_i = 0]$. Ideally one would like all these four local quantities are simultaneously as small as possible for all local sensors, which is impossible. As mentioned in [75], one useful formulation is to find a distributed sequential procedure $D = (T_i, \delta_i)_{i=1}^N$ that (asymptotically) minimizes

$$\max_{i=1,2,\cdots,N} \mathbf{E}_1[T_i] \tag{2.1}$$

subject to the local false alarm constraints:

$$\mathbf{P}_0[\delta_i = 1] \leq \alpha \text{ and } \mathbf{P}_1[\delta_i = 0] \leq \beta \tag{2.2}$$

for all $i = 1, 2, \cdots, N$, where $0 < \alpha, \beta \leq 1/2$ are the pre-specified false alarm bounds.

Note that the objective function in (2.1) can also be replaced by other functions if one wants. For instance, the local criterion $\mathbf{E}_1[T_i]$ can be replaced by $\mathbf{E}_0[T_i]$, or more generally, the Bayesian-type criterion $\pi_0\mathbf{E}_0[T_i] + (1-\pi_0)\mathbf{E}_1[T_i]$. Fortunately, in the context of sequential tests, Wald's (optimal centralized) SPRT can minimize each and every of these criteria, and thus we consider the criterion of $\mathbf{E}_1[T_i]$ here. Moreover, we can also consider other kind of criteria at the global level such as $\min_{i=1,2,\cdots,N} \mathbf{E}_1[T_i]$. Here we do not discuss the appropriateness of different formulations or the corresponding optimality theories, and our focus is to investigate the performance of a specific distributed sequential procedure. For that reason, our results below deal with the local expected sample sizes $\mathbf{E}_1[T_i]$'s themselves, since it is straightforward to extend these local results to the global level such as that in (2.1).

2.2.2 SPRT and CISPRT

Let us first consider the centralized setup when the graph of the network neighborhood structure is complete in the sense that at each time step each local sensor has access to all raw observations over the graph. This is equivalent to the scenario with the fusion center, as each and every local sensor can be regarded as the fusion center. In such scenario, Wald's SPRT is the optimal centralized sequential test under the formulation of (2.1) and (2.2). To define the SPRT, denote the local log-likelihood ratio of the i-th sensor at time step t by

$$\eta_i(t) = \log \frac{f_1(y_i(t))}{f_0(y_i(t))} = \frac{2\mu}{\sigma^2} y_i(t), \tag{2.3}$$

and denote the centralized likelihood ratio statistic up to time t by

$$S_c(t) = \sum_{s=1}^{t}\sum_{i=1}^{N} \eta_i(s) = S_c(t-1) + \sum_{i=1}^{N} \eta_i(t) \tag{2.4}$$

for all $t \geq 1$. The centralized SPRT is then defined by the stopping time

$$T_c = \inf\{t : S_c(t) \notin [\gamma_c^l, \gamma_c^h]\}, \tag{2.5}$$

where γ_c^l and γ_c^h are two pre-specified constants so as to satisfy the false alarm constraints in (2.2). In particular, a commonly used though slightly conservative choice is

$$\gamma_c^l = \log \frac{1-\beta}{\alpha} \quad \text{and} \quad \gamma_c^h = \log \frac{1-\alpha}{\beta}. \tag{2.6}$$

Moreover, since $y_i(t) \sim N(-\mu, \sigma^2)$ under H_0 or $N(\mu, \sigma^2)$ under H_1, the Kullback-Leibler divergence at each local sensor is

$$m = \mathbf{E}_1(\eta_i(t)) = \frac{2\mu^2}{\sigma^2} \tag{2.7}$$

and thus the centralized Kullback-Leibler divergence of the joint observation $\mathbf{Y}(t) = (Y_1(t), \cdots, Y_N(t))$ is Nm. Furthermore, as shown in Wald [91], subject to the false alarm constraint in (2.2), for any sequential test T, distributed or centralized, $\mathbf{E}_1(T) \geq \mathcal{M}(\alpha, \beta)$, where the universal lower bound is given by

$$\mathcal{M}(\alpha, \beta) = \frac{1}{Nm}\Big[(1-\beta)\log\frac{1-\beta}{\alpha} + \beta \log \frac{\beta}{1-\alpha}\Big]. \tag{2.8}$$

Also the centralized SPRT T_c in (2.5) attains this lower bound asymptotically for fixed N and m as $\alpha, \beta \to 0$.

Now let us switch to the distributed setup for a general neighborhood structure where each local sensor can only communicate with its neighborhood sensors. In [75], the authors proposed an interesting CISPRT algorithm where each local sensor makes a local decision based on the weighted average of the local likelihood ratio statistics from itself and its neighborhood sensors. Specifically, at time step t, each i-th local sensor computes its local

test statistic recursively:

$$\begin{aligned} S_i(t) &= w_{ii}S_i(t-1) + \sum_{j\in\Omega_i} w_{ij}S_j(t-1) \\ &\quad + w_{ii}\eta_i(t) + \sum_{j\in\Omega_i} w_{ij}\eta_j(t), \end{aligned} \tag{2.9}$$

for $t = 1, 2, \cdots$, where the initial value $S_i(0) = 0$ and Ω_i is the (one-hop) neighborhood of the i-th sensor. Here the w_{ij}'s are pre-specified weights satisfying

$$w_{ij} \geq 0,\ w_{ii} + \sum_{j\in\Omega_i} w_{ij} = 1,\ \forall i, j \tag{2.10}$$

and the discussion on the choices of the weights w_{ij}'s will be postponed a little bit.

Under the matrix notation, let us collect the weights w_{ij}'s in an $N \times N$ matrix $\mathbf{W}$, where $w_{ij} = 0$ if $(i, j) \notin E$. Denote by $\mathbf{S}(t)$ and $\boldsymbol{\eta}(t)$ as the $N \times 1$ vectors $(S_1(t), \cdots, S_N(t))^T$ and $(\eta_1(t), \cdots, \eta_N(t))^T$. The local test statistics can be updated recursively as

$$\mathbf{S}(t) = \mathbf{W}\Big(\mathbf{S}(t-1) + \boldsymbol{\eta}(t)\Big). \tag{2.11}$$

for $t \geq 1$.

For the CISPRT, each i-th sensor makes a local decision at time

$$T_i = \inf\{t \geq 1 : S_i(t) \notin [\gamma_i^l, \gamma_i^h]\}, \tag{2.12}$$

for some pre-specific thresholds $\gamma_i^l < 0 < \gamma_i^h$. When stopping, the i-th sensor makes a local decision

$$\delta_i = \begin{cases} 0, & \text{if } S_i(T_i) \leq \gamma_i^l; \\ 1, & \text{if } S_i(T_i) \geq \gamma_i^h. \end{cases} \tag{2.13}$$

From the pure mathematical viewpoint, the stopping times T_i's in (2.12) depend on the

properties of $S_i(t)$'s which are a component of the N-dimensional random walks $\mathbf{S}(t)$'s. One may be able to apply the classical renewal theory to analyze the "asymptotic" properties of the stopping times T_i's in (2.12), but unfortunately such an approach will involve "constant" terms for overshoot analysis that are exponentially increasing as the dimension N increases. In particular, when the number N of sensors is large, such constant terms can be huge, and thus the corresponding asymptotic analysis can be meaningless under any reasonable practical setting of distributed detection. Here we provide an alternative approach that yields the same first-order asymptotic result as in the classical renewal theory when the number N of sensors is 1 or very small, but has a potential to derive useful oracle properties of stopping times under the setting of high dimension N for Gaussian data.

2.2.3 Spectral Graph Theory and Weight Matrix $\mathbf{W}$

In this subsection, let us present some basic materials for spectral graph theory that are related to the main assumption for our network structure and the design of the weight matrix $\mathbf{W}$ in (2.10). Also see [61] for a more complete introduction of graph theory.

Recall our network neighbor structure is characterized by the undirected simple graph $G = (V, E)$. In spectral graph theory, the degree matrix $\mathbf{D}$ is an $N \times N$ diagonal matrix with the i-th diagonal being d_i, the degree of the i-th vertex. The adjacency matrix $\mathbf{A}$ is a $(0, 1)$-matrix with zeros on its diagonal and $A_{ij} = 1$ if and only if the i-th vertex and the j-th vertex are connected for all $1 \leq i \neq j \leq N$. The Laplacian matrix is then defined as

$$\mathbf{L} = \mathbf{D} - \mathbf{A}. \tag{2.14}$$

Alternatively, for the Laplacian matrix $\mathbf{L}$, its elements are given by

$$L_{i,j} = \begin{cases} d_i & \text{if } i = j; \\ -1 & \text{if } i\text{-th and } j\text{-th vertex connected for } i \neq j; \\ 0 & \text{otherwise.} \end{cases}$$

The Laplacian matrix $\mathbf{L}$ is a positive semidefinite matrix, and thus has N non-negative eigenvalues:

$$0 = \lambda_1(\mathbf{L}) \leq \lambda_2(\mathbf{L}) \leq \cdots \leq \lambda_N(\mathbf{L}). \tag{2.15}$$

Moreover, the number of times 0 appears as an eigenvalue of the Laplacian matrix $\mathbf{L}$ is the number of connected components in the graph. Equivalently, a graph is connected if and only if $\lambda_2(\mathbf{L}) > 0$, see [10].

Our main assumption on the network neighborhood structure is as follows.

Assumption 2.1. *The graph $G = (V, E)$ is connected, or equivalently, the second smallest eigenvalue of the Laplacian matrix $\mathbf{L}$ is positive, i.e., $\lambda_2(\mathbf{L}) > 0$.*

Next, let us discuss the choices of the weight matrix $\mathbf{W}$ in (2.11). From the technical or algorithm viewpoint, the weight matrix $\mathbf{W}$ can be arbitrary as long as $\mathbf{W}$ is a stochastic matrix in the sense of satisfying (2.10). However, in the context of distributed sequential detection, the implicit assumption is that $w_{ij} > 0$ if and only if the i-th and j-th sensors are neighbors. There are still many reasonable choices for the weight matrix $\mathbf{W}$, and one useful one is to define

$$\mathbf{W} = \mathbf{I}_{N\times N} - \delta\mathbf{L}, \tag{2.16}$$

where $\mathbf{I}_{N\times N}$ is the $N \times N$ identity matrix and δ is a small positive constant so that all elements of $\mathbf{W}$ are positive and thus (2.10) holds. Under this choice of the weight matrix $\mathbf{W}$, for a given i-th sensor/vertex, it assigns a small but equal weight of $w_{ij} = \delta$ to all of its d_i neighbor sensors, and assigns most weight $w_{ii} = 1 - \delta d_i$ to itself.

An interesting fact of the weight matrix $\mathbf{W}$ in (2.16) is that it is symmetric ($w_{ij} = w_{ji}$) and irreducible, and the latter is due to the fact that the graph is connected under Assumption 2.1. In addition, it is straightforward to show that as a stochastic matrix sat-

isfying (2.10), the matrix $\mathbf{W}$ has the largest eigenvalue 1, and the second largest eigenvalue, denote by r, is strictly less than 1. Recall that any $N \times N$ symmetric matrix can be written as $\mathbf{W} = \sum_{i=1}^{N} \lambda_i u_i v_i^T$, where u_i and v_i are singular vectors associated with the i-th largest eigenvalue λ_i. For the stochastic symmetric matrix $\mathbf{W}$ satisfying (2.10), we have $\lambda_1 = 1$ and $u_1 = v_1 = \frac{1}{\sqrt{N}}\mathbf{1}$, where $\mathbf{1}$ is an N-dimensional all one vector. Thus $\mathbf{W} - \mathbf{J} = \sum_{i=2}^{N} \lambda_i u_i v_i^T$, where $\mathbf{J} = \frac{1}{N}\mathbf{1}\mathbf{1}^T$ be an $N \times N$ matrix with all entries being the constant $1/N$. This eigenvalue decomposition representation shows that the largest eigenvalue of $\mathbf{W} - \mathbf{J}$ is simply the second largest eigenvalue of $\mathbf{W}$. Hence, under our notation, the second largest eigenvalue r of $\mathbf{W}$ can be characterized by the spectral norm (or the largest eigenvalue) of the matrix $\mathbf{W} - \mathbf{J}$, i.e.,

$$r = ||\mathbf{W} - \mathbf{J}||. \tag{2.17}$$

2.3 Improved Properties of CISPRT

In this section, we derive our main theoretical properties of the CISPRT procedure in (2.12) and (2.13) under Assumption 2.1 when the network is connected. Note that there are two thresholds, γ_i^l and γ_i^h, in the CISPRT procedure in (2.12). At the high level, the upper bound γ_i^h is closely related to Type I error probability and the expected sample size under H_1, whereas the lower bound γ_i^l is closely related to Type II error probability and the expected sample size under H_0. For simplicity of the technical proofs, below our main theorem will focus on the effects of the upper bound γ_i^h on the Type I error probability and the expected sample size under H_1 of the CISPRT. The usefulness of the our main theorem is illustrated in several corollaries which consider the symmetric scenario when the lower and upper bounds satisfy $\gamma_i^l = -\gamma_i^h$.

Let us first summarize the main theoretical results of [75] that are closely related to our

paper. For a given weight matrix $\mathbf{W}$, denote by

$$\rho = 1 - \exp\left(-\frac{Nm}{4(Nr^2+1)}\right), \tag{2.18}$$

where m is the Kullback-Leibler divergence in (2.7) and r is the second largest eigenvalue of $\mathbf{W}$ and can be rewritten as in (2.17). It was shown in [75] that the Type I error of the CISPRT algorithm satisfies

$$\mathbb{P}_0(\delta_i = 1) \leq 2\rho^{-1} \exp\left(-\frac{7}{8}\frac{N}{Nr^2+1}\gamma_i^h\right), \tag{2.19}$$

and its expected sample size satisfies

$$\begin{aligned} & \frac{1}{m}[\gamma_i^h - \mathbb{P}_1(\delta_i = 0)(\gamma_i^h - \gamma_i^l) - c] \\ \leq \quad & \boldsymbol{E}_1(T_i) \leq \frac{5}{4}\frac{1}{m}\gamma_i^h + \rho^{-1} + 1, \end{aligned} \tag{2.20}$$

see Theorem 4.1, Theorem 4.7, and equation (49) of [75]. Here the constant $c > 0$ in the lower bound in (2.20) is independent of the thresholds γ_i^h, γ_i^l and is a complicated function of the network topology and the Gaussian model statistics, see equations (47)-(49) of [75]. Note that the original upper bound does not have the constant 1 in the right-hand side of (2.20), but we found out that the original proof in [75] contains a minor mistake to count the number of integers in the interval $0 \leq t \ \leq a$ as a, not $a+1$. Thus we add 1 here so that the results are mathematically rigorous.

Moreover, the authors in [75] considered the asymptotic properties of the symmetric CISPRT with $\gamma_i^l = -\gamma_i^h$ subject to the local false alarm constraints with $\alpha = \beta = \epsilon$ as $\epsilon \to 0$. By (2.19), a conservative choice of the thresholds $\gamma_i^h = -\gamma_i^h$ is

$$\gamma_i^{h,0} = -\gamma_i^h = \frac{8(Nr^2+1)}{7N}\left(\log(2\rho^{-1}) + \log\epsilon^{-1}\right), \tag{2.21}$$

see equation (19) in Theorem 4.1 of [75]. For the CISPRT with the thresholds in (2.26), the authors in [75] then compared its expected stopping time with the universal lower bound in (2.8):

$$1 \leq \limsup_{\epsilon \to 0} \frac{\mathbf{E}_1(T_i)}{\mathcal{M}(\epsilon)} \leq \frac{10}{7}(Nr^2+1), \tag{2.22}$$

where the lower bound is trivial since $\mathcal{M}(\epsilon) = \mathcal{M}(\alpha = \epsilon, \beta = \epsilon)$ is the universal lower bound in (2.8) for any stopping times satisfying the local false alarm constraints when $\alpha = \beta = \epsilon$, and the factor $\frac{10}{7}$ in the upper bound are based on the factors $\frac{8}{7} \times \frac{5}{4}$ from (2.26) and (2.20).

In this paper, we improve the constants $7/8, 5/4$ and $10/7$ of the upper bounds in (2.19)-(2.22) in the original CISPRT paper [75] to $1, 1$, and 1, respectively. The price we pay is to add additional terms that can be thought of as the second-order terms in the asymptotic setting.

Our main new results can be summarized in the following theorem and corollaries.

Theorem 2.1. *For the CISPRT, at any given i-th local sensor, the Type I error probability satisfies*

$$\begin{aligned} \mathbb{P}_0(\delta_i = 1) \quad &\leq \quad \rho^{-1} \exp\left\{ -\frac{N}{Nr^2+1}\gamma_i^h + \right. \\ &\qquad \left. + \log\left(1 + \frac{N}{4(Nr^2+1)}\gamma_i^h\right) + 1 \right\}, \end{aligned} \tag{2.23}$$

and its expected sample size under H_1 satisfies

$$\boldsymbol{E}_1(T_i) \leq \frac{1}{m}\gamma_i^h + \frac{1}{2}\rho^{-1/2}\left(\sqrt{\frac{1}{m}\gamma_i^h} + 1\right) + 1. \tag{2.24}$$

Here ρ is a constant in (2.18), and m is the Kullback-Leibler divergence in (2.7).

Theorem 2.1 deals with the non-asymptotic properties of the CISPRT, and holds for any network structure regardless of how large the number N of sensors is. The proof of

Theorem 2.1 is very technical and will be postponed to Section 2.5. Here let us present some corollaries to illustrate the usefulness of our non-asymptotic upper bounds in (2.23) and (2.24), which turn out to be first-order tight as the threshold γ_i^h goes to ∞ when the network structures are fixed.

Corollary 2.1. *For the symmetric CISPRT with $\gamma_i^l = -\gamma_i^h$, as the threshold $\gamma_i^h \to \infty$, we have $\mathbb{P}_0(\delta_i = 1) = \mathbb{P}_1(\delta_i = 0) \to 0$, and*

$$\boldsymbol{E}_1(T_i) = (1 + o(1))\frac{1}{m}\gamma_i^h. \tag{2.25}$$

Thus the upper bound (2.24) on the expected sample size is asymptotically accurate up to first-order.

Corollary 2.2. *Under the symmetric local false alarm constraints $\alpha = \beta = \epsilon$ in (2.2), consider the CISPRT algorithm with the thresholds $\gamma_i^h = -\gamma_i^h$ defined as*

$$\gamma_i^h = -\gamma_i^l = \frac{Nr^2 + 1}{N}\Delta_\epsilon, \tag{2.26}$$

where Δ_ϵ is the solution Δ of

$$\Delta - \log(1 + \frac{1}{4}\Delta) - 1 = \log(\rho^{-1}) + \log(\epsilon^{-1}). \tag{2.27}$$

Then relation (2.23) implies that this CISPRT algorithm satisfies local false alarm constraints $\alpha = \beta = \epsilon$ in (2.2). Moreover, by (2.25), as $\alpha = \beta = \epsilon \to 0$,

$$\lim_{\epsilon \to 0} \frac{\mathbf{E}_1(T_i)}{\mathcal{M}(\epsilon)} = Nr^2 + 1 \tag{2.28}$$

for all $i = 1, 2, \cdots, N$.

Corollary 2.3. *Among the family of the CISPRTs with the weight matrix $\mathbf{W}$ of the form in (2.16), the optimal one with the smallest asymptotically expected sample size in (2.28) is*

attained by the weight matrix $\mathbf{W}_{opt}$ *in (2.16) with*

$$\delta_{opt} = 2/(\lambda_N(\mathbf{L}) + \lambda_2(\mathbf{L})), \tag{2.29}$$

which yields the minimum r *value*

$$r_{opt} = \frac{\lambda_N(\mathbf{L}) - \lambda_2(\mathbf{L})}{\lambda_N(\mathbf{L}) + \lambda_2(\mathbf{L})}, \tag{2.30}$$

where $\lambda_N(\mathbf{L})$ *and* $\lambda_2(\mathbf{L})$ *are the largest and second smallest eigenvalues in (2.15) for the Laplacian matrix* $\mathbf{L}$. *In particular, for the complete graph when each sensor is connected to all other sensors, we have* $\delta_{opt} = 1/N$ *and* $r_{opt} = 0$. *In such a case, the CISPRT with the optimal weight matrix is equivalent to the centralized SPRT* T_c *in (2.5), and relation (2.28) is consistent with the well-known fact that the centralized SPRT* T_c *in (2.5) attains the universal lower bound* $\mathcal{M}(\epsilon)$ *asymptotically as* $\alpha = \beta = \epsilon \to 0$.

Corollary 2.4. *The threshold in (2.26) is first-order asymptotically tight in the sense that if the thresholds* $\gamma_i^h = -\gamma_i^l = M_\epsilon \frac{Nr^2+1}{N} \Delta_\epsilon$ *with* $\liminf_{\epsilon\to 0} M_\epsilon = M < 1$ *being a constant that does not depend on* N *and* r, *then the corresponding CISPRT algorithm cannot satisfy the local false alarm constraints* $\alpha = \beta = \epsilon$ *in (2.2). That is, the upper bound (2.23) on Type I error probability is asymptotically accurate up to first-order in the logarithm scale to derive the thresholds.*

The proofs of these corollaries follow directly from Theorem 2.1 and other well-known facts, and below we present a high-level sketch of the proofs.

Proof of Corollary 2.1: On the one hand, the upper bound (2.24) in Theorem 2.1 implies that $\boldsymbol{E}_1(T_i) \leq (1 + o(1))\frac{1}{m}\gamma_i^h$ as $\gamma_i^h \to \infty$. On the other hand, by the symmetric arguments, we have $\mathbb{P}_1(\delta_i = 0) = \mathbb{P}_0(\delta_i = 1)$, which goes to 0 as $\gamma_i^h \to \infty$ by the upper bound (2.23). Then the lower bound in (2.20) implies that $\boldsymbol{E}_1(T_i) \geq (1 + o(1))\frac{1}{m}\gamma_i^h$. Thus relation (2.25) holds.

Proof of Corollary 2.2: Equating the upper bound (2.23) to ϵ and solving it to yield the desired threshold h_i^h in (2.26) and (2.27). As $\epsilon \to 0$, we have $\Delta_\epsilon \sim \log(\epsilon^{-1})$ by (2.27) and $\mathcal{M}(\epsilon) \sim \frac{1}{Nm}\log(\epsilon^{-1})$ by (2.8). Here and below we denote by $x(\epsilon) \sim y(\epsilon)$ if and only if $\lim_{\epsilon\to 0}(x(\epsilon)/y(\epsilon)) = 1$. Relation (2.28) then follows directly from (2.25) and (2.26).

Proof of Corollary 2.3: By (2.28), it suffices to minimize the second largest eigenvalue r of $\mathbf{W}$ among all weight matrices $\mathbf{W}$ of the form in (2.16). It is shown in reference [95] that the corresponding optimal solution and value are given by (2.29) and (2.30).

As for the complete graph, its Laplacian matrix is $\mathbf{L} = N\mathbf{I}_{N\times N} - \mathbf{1}\mathbf{1}^T$, where the vector $\mathbf{1}$ is an N-dimensional all one vector. An elementary algebra shows that the eigenvalues of $\mathbf{L}$ are $\lambda_1 = 0$ and $\lambda_2 = \lambda_3 = \cdots = \lambda_N = N$. By (2.29) and (2.30), we have $\delta_{opt} = 1/N$ and $r_{opt} = 0$. Hence, for the CISPRT with the optimal weight matrix under the complete graph scenario, each local sensor is to put equal weights to all raw sensor observations, and each and every local sensor essentially runs the optimal centralized SPRT in (2.4).

Proof of Corollary 2.4: let us prove by contradiction, and assume the CISPRT algorithm with those choices of $\gamma_i^h = -\gamma_i^l$ would satisfy the local false alarm constraints in (2.2). By (2.25), the relation (2.28) would become $\liminf_{\epsilon\to 0}\frac{\mathbf{E}_1(T_i)}{\mathcal{M}(\epsilon)} = M(Nr^2+1)$ for any network structures. By Corollary 2.3, a special case is the complete graph with $r = 0$, which would imply that $\liminf_{\epsilon\to 0}\frac{\mathbf{E}_1(T_i)}{\mathcal{M}(\epsilon)} = M < 1$. In other words, we would have $\mathbf{E}_1(T_i) < \mathcal{M}(\epsilon)$. This is a contradiction since $\mathcal{M}(\epsilon)$ is the universal lower bounds for any stopping times satisfying the local false alarm constraints $\alpha = \beta = \epsilon$ in (2.2). Thus the corollary holds.

2.4 Simulation Studies

In this section, we report our simulation study results to illustrate the usefulness of our improved performance properties of CISPRT algorithm.

We use random graph to generate the neighborhood structure of sensor as follow. Assume the N sensors correspond to N random points in a unit square $[0,1]\times[0,1]$. Two

sensors are connected if and only if the distance of the corresponding two points is less than the connectivity parameter g. In our simulation studies, we consider two choices of $N = 30$ and 300, and two choices of the connectivity parameter $g = 0.3$ and 0.9. In other words, we consider a total of $2 \times 2 = 4$ different network structures. For each of these four given networks, the raw sensor observations $y_i(t)$'s are assumed to be $N(\mu_1, \sigma^2)$ under the alternative hypothesis H_1 and $N(-\mu_1, \sigma^2)$ under the null hypothesis H_0. Here we set the mean $\mu_1 = 1/\sqrt{2}$ and $\sigma = 1$, so that the local Kullback-Leibler divergence in (2.7) becomes $m = 2\mu^2/\sigma^2 = 1$ at each local sensor.

Our focus is on the performance of the CISPRT in (2.13) in each of these four given networks, as the Type I and Type II error probabilities constraints $\alpha = \beta = \epsilon$ vary from 10^{-8} to 10^{-4} with step size 10^{-6}. For simplicity, we consider the symmetric scenario when the lower and upper bounds γ_i^l and γ_i^h of the CISPRT in (2.13) are given by $\gamma_i^l \equiv -\gamma$ and $\gamma_i^h \equiv \gamma$ for all $i = 1, \cdots, N$ for some $\gamma > 0$. In our simulation studies, we choose γ based on (2.26) and (2.27). For each give threshold γ, we simulate the expected sample size of the CISPRT in (2.13) under H_1 based on $M = 2000$ Monte Carlo runs.

For the purpose of easy understanding, Figure 2.1 compares our improved bound (red solid line) in (2.28) with three other estimates of $\mathbf{E}_1(T_i)/\mathcal{M}(\epsilon)$ of the CISPRT: (i) Sahu and Kar's upper bound (blue dashed line) in (2.22); (ii) the Monte Carlo simulated expected sample size, $\mathbf{E}_1(T_i)/\mathcal{M}(\epsilon)$ (purple dotted line); and (iii) the first term in (2.20) of the lower bound divided by (2.8), $\frac{(1-2\epsilon)\gamma_i^h/m}{(1-2\epsilon)\log(1-\epsilon/\epsilon)/(Nm)} = \gamma N/\log(1-\epsilon/\epsilon)$, (green dotdash line). From the plots, our bounds on the ratio of expected sample size are much closer to the Monte Carlo simulated ratio than Sahu and Kar's original upper bound, no matter the number of sensors and the connectivity parameter.

It is interesting to see the simulated $\mathbf{E}_1(T_i)/\mathcal{M}(\epsilon)$ falls between our improved bound and the lower bound. However, note that the actual lower bound in (2.20) includes an uncomputable constant, and thus it is not as close to the Monte Carlo simulated $\mathbf{E}_1(T_i)/\mathcal{M}(\epsilon)$ as Figure 2.1 illustrates. Also note that the first order term of our improved upper bound in

(2.24) is same as that of the lower bound as ϵ goes to 0, and this confirms that our improved bound in (2.28) is attainable and our improved upper bound of $\mathbf{E}_1(T_i)$ in (2.24) is sharp up to first-order.

2.5 Proof of Theorem 2.1

This section is devoted to prove Theorem 2.1 under the non-asymptotic setting. Let us fix the i-th local sensor, and investigate the properties of the stopping time T_i in (2.12) of the CISPRT at this specific local sensor that are related to the upper bound γ_i^h. Since the detailed proof is technical and involves many subscripts, we decide to abuse the notation and denote the stopping time T_i and the upper bound γ_i^h in (2.12) simply by T and γ.

Let us first provide the high-level idea to prove Theorem 2.1. Note that the Type I error probability can be written as

$$\mathbb{P}_0(\delta_i = 1) = \mathbb{P}_0(S_i(T) \geq \gamma), \tag{2.31}$$

where $S_i(T)$ is the value of the local test statistic in (2.9) at the stopping time T. Note that in the classical sequential analysis for the centralized setting, it is standard to use the change of measures arguments, and rewrite the Type I error probability as its equivalent form of $\mathbf{E}_1(e^{-S_c(T)}I(S(T) \geq \gamma))$, where $S_c(T)$ is the centralized log-likelihood ratio in (2.4). The analysis on this error probability analysis and the expected sample size $\mathbf{E}_1(T)$ is then based on the renewal theory and overshoot analysis for random walks over time t. Unfortunately, such approach breaks down for distributed setting when the centralized test statistic $S_c(t)$ and the local test statistic $S(t)$ can be completely different. Moreover, with large number N sensors, the overshoot analysis often involves constants that are exponentially increasing as N increases and thus the corresponding analysis can be meaningless under the practical setting.

Sahu and Kar [75] proposed an alternative method to bound the Type I error probability

and expected sample size directly. Specifically, note that

$$\begin{aligned}
\mathbb{P}_0(\delta_i = 1) &= \mathbb{P}_0(S_i(T) \geq \gamma) \\
&= \sum_{t=1}^{\infty} \mathbb{P}_0(T = t, S_i(t) \geq \gamma) \\
&\leq \sum_{t=1}^{\infty} \mathbb{P}_0(S_i(t) \geq \gamma) \\
&= \sum_{t=1}^{\infty} \mathbb{Q}\left(\frac{\gamma - \mu_0^*(t)}{\sqrt{V_0^*(t)}}\right). \qquad (2.32)
\end{aligned}$$

Here $\mathbb{Q}(u) = \boldsymbol{P}(N(0,1) > u)$, and the local test statistics $S_i(t)$ are Gaussian distributed under H_0, say, $N(\mu_0^*(t), V_0^*(t))$, at any fixed t, since the raw sensor observations are Gaussian. Meanwhile, the expected sample size, $\boldsymbol{E}_1(T)$, is bounded by

$$\begin{aligned}
\boldsymbol{E}_1(T) &= \sum_{t=0}^{\infty} \boldsymbol{P}_1(T > t) \\
&\leq \sum_{t=0}^{\infty} \boldsymbol{P}_1(S_i(t) < \gamma) \\
&\leq \sum_{t=0}^{\infty} \mathbb{Q}\left(\frac{\gamma - \mu_1^*(t)}{\sqrt{V_1^*(t)}}\right), \qquad (2.33)
\end{aligned}$$

where the local test statistics $S_i(t)$ are Gaussian $N(\mu_1^*(t), V_1^*(t))$ under H_1.

Due to the similarity between (2.32) and (2.33), below we will focus on the Type I error probability analysis in (2.32). By (2.7) and (2.9), it was showed in [75] that

$$\mu_0^*(t) = -mt \qquad \text{and} \qquad V_0^*(t) \leq \frac{2m(Nr^2+1)t}{N}. \qquad (2.34)$$

In [75], the authors then combined these above results together to bound the infinite sum in (2.32) by splitting the interval $t \in [1, \infty)$ into four subintervals:

$$[1, \frac{\gamma}{2m}], \quad (\frac{\gamma}{2m}, \frac{\gamma}{m}], \quad (\frac{\gamma}{m}, \frac{2\gamma}{m}], \quad (\frac{2\gamma}{m}, \infty). \qquad (2.35)$$

Bounding the sum in each of these four subintervals leads to the result in (2.19) in the original CISPRT paper.

The direct approach in (2.32) - (2.35) is non-asymptotic, but unfortunately it is too crude in general. Indeed, if we apply them directly to investigate the Type I error probability or expected sample sizes of the centralized SPRT, the results will be much looser as compared with those from the classical renewal theory: while the first-order terms have the same order, the coefficients from the direct approach in (2.32) - (2.35) are much larger.

After a careful analysis, we find out that the main reason is caused by the middle two subintervals in (2.35), and the direct approach in (2.32) - (2.35) can be refined to provide better bounds if we further split each of the middle two subintervals into k sub-subintervals, for some suitable choice of k that will be optimally determined later. In fact, when we applied this new refined approach to investigate the Type I error probability or expected sample sizes of the centralized SPRT, then the corresponding results are first-order asymptotically equivalent to those from the classical renewal theory. This suggests that the refined direct approach yields an accurate upper bound for complete graph regardless of the number N of sensors, and thus may also lead to good bounds for other network structures. We acknowledge that the proof techniques are essentially that of [75], except for the new found techniques as far as splitting the intervals are concerned in (2.36) below.

Now we are ready to provide the detailed, rigorous proof of (2.23). First, we further split each of the middle two subintervals into k sub-subintervals as follows. Let $\ell = \frac{\gamma}{2m}$ or $\frac{\gamma}{m}$, and we propose to further split the subinterval $[\ell, 2\ell]$ as k sub-subintervals:

$$(\frac{k+j-1}{k}\ell,\ \frac{k+j}{k}\ell] \qquad \text{for } j = 1, \cdots, k. \tag{2.36}$$

Relation (2.23) in Theorem 2.1 can be proved by bounding the infinite sum in (2.32) through these subintervals.

Second, we will use heavily the following well-known fact for $N(0,1)$ distribution:

$$\mathbb{Q}(x) = \mathbf{P}(N(0,1) > x) \le \frac{1}{2}\exp\Big(-\frac{x^2}{2}\Big) \text{ for all } x > 0. \tag{2.37}$$

Also $Q(x)$ is decreasing as a function of x, and thus replacing $V_0^*(t)$ by its upper bound in (2.34) yields an upper bound of (2.32).

Next, by (2.32), (2.34) and (2.37), we have

$$\begin{aligned}
\mathbb{P}_0(\delta_i = 1) &\le \sum_{t=1}^{\infty} \mathbb{Q}\left(\frac{\gamma + mt}{\sqrt{\frac{2mt(Nr^2+1)}{N}}}\right) \\
&\le \frac{1}{2}\sum_{t=1}^{\infty} \exp\left\{\frac{-N(\gamma+mt)^2}{4mt(Nr^2+1)}\right\} \\
&= \frac{1}{2}(A_1 + A_2 + A_3 + A_4) \\
&= \frac{1}{2}(A_1 + \sum_{j=1}^{k} A_{2j} + \sum_{j=1}^{k} A_{3j} + A_4),
\end{aligned} \tag{2.38}$$

where A_1, A_2, A_3 and A_4 denote the corresponding sum when the integer index t ranges over the subintervals in (2.35). Here A_{2j} and A_{3j} are defined as the summation over the sub-subintervals in (2.36) for $\ell = \frac{\gamma}{2m}$ or $\frac{\gamma}{m}$. In other words,

$$A_{2j} = \sum_{t=\lfloor \frac{k+j-1}{k}\ell \rfloor + 1}^{\lfloor \frac{k+j}{k}\ell \rfloor} \exp\left\{\frac{-N(\gamma+mt)^2}{4mt(Nr^2+1)}\right\}$$

with $\ell = \frac{\gamma}{2m}$, and A_{3j} is defined similarly with $\ell = \frac{\gamma}{m}$. Here and below $\lfloor x \rfloor$ denotes the largest integer $\le x$.

Sahu and Kar [75] proved their results by bounding A_1, A_2, A_3 and A_4, and here we refine their results by bounding A_{2j}'s and A_{3j}'s. The main mathematical tool is the simple

fact that when $a \leq t \leq b$, for $c = N/(4(Nr^2+1))$,

$$\begin{aligned} & \sum_{t=a}^{b} \exp\left\{-c\frac{(\gamma+mt)^2}{mt}\right\} \\ \leq & \sum_{t=a}^{b} \exp\left\{-c\big(\frac{\gamma^2}{mb}+2\gamma+mt\big)\right\} \\ = & \exp\left\{-c\big(\frac{\gamma^2}{mb}+2\gamma\big)\right\}\frac{\exp(-cma)-\exp(-cm(b+1))}{1-\exp(-cm)} \\ \leq & \rho^{-1}\exp\left\{-c\big(\frac{\gamma^2}{mb}+2\gamma\big)\right\}\exp(-cma), \end{aligned} \tag{2.39}$$

where the constant $\rho = 1-\exp(-cm)$ is defined in (2.18).

In order to help casual readers better understand our main ideas, let us first provide the bounds of the original CISPRT paper [75] on A_1 and A_2. Applying (2.39) to the case of $a = 1$ and $b = \gamma/(2m)$, we have

$$\begin{aligned} A_1 &\leq \rho^{-1}\exp(-4c\gamma)\exp(-cm) \\ &\leq \rho^{-1}\exp\Big(-\frac{N\gamma}{Nr^2+1}\Big). \end{aligned} \tag{2.40}$$

where the last relations follows from the fact that $\exp(-cm) < 1$ and the definition of $c = N/(4(Nr^2+1))$.

Similarly, applying (2.39) to the case of $a = \frac{\gamma}{2m}$ and $b = \frac{\gamma}{m}$, we have

$$\begin{aligned} A_2 &\leq \rho^{-1}\exp(-3c\gamma)\exp(-\frac{1}{2}c\gamma) \\ &= \rho^{-1}\exp(-\frac{7}{2}c\gamma) \\ &= \rho^{-1}\exp\Big(-\frac{7}{8}\frac{N\gamma}{Nr^2+1}\Big). \end{aligned} \tag{2.41}$$

It is easy to see that A_3 satisfies (2.41), whereas A_4 satisfies (2.40). A combination of (2.38) with the bounds in (2.40) and (2.41) yields (2.19), which is the upper bound of $\mathbb{P}_0(\delta_i = 1)$ derived in [75].

To improve the upper bound in (2.19) of [75], our key observation is that the bound

in (2.41) for A_2 and A_3 can be further reduced. For that purpose, let us consider the A_{2j} over the j-th sub-subinterval in (2.36), and apply (2.39) to the case of $a = \frac{k+j-1}{k}\frac{\gamma}{2m}$ and $b = \frac{k+j}{k}\frac{\gamma}{2m}$. Then for $j = 1, 2, \cdots, k$, we have

$$\begin{aligned} A_{2j} &\leq \rho^{-1} \exp\Big(-2c\gamma \frac{2k+j}{k+j} \Big) \exp\Big(-c\gamma \frac{k+j-1}{2k} \Big) \\ &= \rho^{-1} \exp(-2c\gamma) \exp\Big(-c\gamma (\frac{2k}{k+j} + \frac{k+j-1}{2k}) \Big) \\ &\leq \rho^{-1} \exp(-2c\gamma) \exp\Big(-c\gamma (2 - \frac{1}{2k}) \Big) \\ &= \rho^{-1} \exp\left\{ -\frac{8k-1}{8k} \frac{N\gamma}{Nr^2+1} \right\}, \end{aligned} \tag{2.42}$$

where the second to last relation follows from the simple fact that $u + 1/u \geq 2$ for $u = 2k/(k+j)$, and the last equation is from the definition of $c = N/(4(Nr^2+1))$.

It is useful to discuss the implication of (2.42) and compare the corresponding upper bound on A_2 with those in (2.41). By (2.42), we have

$$A_2 \leq k\rho^{-1} \exp\left\{ -\frac{8k-1}{8k} \frac{N\gamma}{Nr^2+1} \right\} \tag{2.43}$$

It is interesting to see that (2.41) is a special case of (2.43) with $k = 1$. By increasing the value of k or the number of sub-subintervals, we can reduce the factor $-7/8$ in the exponent term of (2.41) to a smaller value that is close to -1, and the price we pay is the multiplication factor k.

Similarly, the same technique of (2.42) is applied to A_{3j} or A_3, and we have

$$A_{3j} \leq \rho^{-1} \exp\left\{ -\frac{4k+3}{4k+4} \frac{N\gamma}{Nr^2+1} \right\}. \tag{2.44}$$

for $j = 1, \cdots, k$. On the one hand, when $k = 1$, the upper bounds in (2.42) and (2.44) are the same with the factor $7/8$ in the exponential term. On the other hand, for a general $k > 1$, the upper bound in (2.44) is larger, which can also be used to bound all $2(k+1)$ terms in A_1, A_{2j} and A_4. By (2.38), the Type I error probability of the CISPRT is bounded

by

$$\begin{aligned}
\mathbb{P}_0(\delta_i = 1) &\leq \frac{1}{2}(A_1 + \sum_{j=1}^{k} A_{2j} + \sum_{j=1}^{k} A_{3j} + A_4) \\
&\leq (k+1)\rho^{-1} \exp\left\{-\frac{4k+3}{4k+4}\frac{N\gamma}{Nr^2+1}\right\} \\
&= \rho^{-1} \exp\left\{-\frac{N\gamma}{Nr^2+1}\right\} \times \\
&\quad \times \exp\left\{\log(u) + \frac{1}{u}\frac{N\gamma}{4(Nr^2+1)}\right\}, \qquad (2.45)
\end{aligned}$$

where $u = k + 1$. A simple calculus analysis shows that for any given $D > 0$, the function $\log(u) + D/u$ is minimized at $u_{opt} = D$ with the minimum value of $\log(D) + 1$. However, a subtly here is that we should restrict u to be integers. The good news is that (2.45) holds for any integer $u = k+1$ and thus we can choose a specific integer $u^* = \lceil D \rceil$, the smallest integer $\geq D$. For this specific choice of u^*, we have

$$\log(u^*) + \frac{D}{u^*} \leq \log(D+1) + \frac{D}{D} = \log(D+1) + 1, \qquad (2.46)$$

which is asymptotically equivalent to the minimum bound $\log(D) + 1$ for large D. Combining the above results together yields (2.23), which completes the proof of Type I error properties of the CISPRT algorithm.

The proof of (2.24) for the expected sample size is similar, except with different subintervals. By (2.7) and (2.9), we can show that

$$\mu_1^*(t) = mt \qquad \text{and} \qquad V_1^*(t) \leq \frac{2m(Nr^2+1)t}{N}. \qquad (2.47)$$

By (2.33), for the CISPRT,

$$\begin{aligned}
\boldsymbol{E}_1(T) &\leq \sum_{t=0}^{\infty} \mathbb{Q}(\frac{tm-\gamma}{\sqrt{2tm(Nr^2+1)/N}}) \\
&= B_1 + B_2 + B_3 + B_4, \qquad (2.48)
\end{aligned}$$

where B_1, B_2, B_3, B_4 correspond to the summation over t in each of the following four subintervals:

$$[0, \frac{\gamma}{m}], \quad (\frac{\gamma}{m}, \frac{3\gamma}{2m}], \quad (\frac{3\gamma}{2m}, \frac{2\gamma}{m}], \quad (\frac{2\gamma}{m}, \infty). \tag{2.49}$$

It turns out that the bounds derived in [75] for B_1, B_3, B_4 are tight, and the bound on B_2 is too loose. To be more specific, in [75], the authors used the similar technique for Type I error to show that $B_3 \leq \frac{1}{2}\rho^{-1}$ and $B_4 \leq \frac{1}{2}\rho^{-1}$, and also bound B_1 and B_2 by

$$\begin{aligned} B_1 &= \sum_{t=0}^{\lfloor \gamma/m \rfloor} \mathbb{Q}(\frac{tm-\gamma}{\sqrt{2tm(Nr^2+1)/N}}) \qquad (2.50) \\ &\leq \sum_{t=0}^{\lfloor \gamma/m \rfloor} 1 \leq \frac{\gamma}{m} + 1, \\ B_2 &= \sum_{t=\lfloor \gamma/m \rfloor + 1}^{\lfloor 3\gamma/(2m) \rfloor} \mathbb{Q}(\frac{tm-\gamma}{\sqrt{2tm(Nr^2+1)/N}}) \\ &\leq \sum_{t=\lfloor \gamma/m \rfloor + 1}^{\lfloor 3\gamma/(2m) \rfloor} \frac{1}{2} = \frac{\gamma}{4m}, \\ B_3 &\leq \frac{1}{2}\rho^{-1}, \quad \text{and} \quad B_4 \leq \frac{1}{2}\rho^{-1} \end{aligned}$$

Here B_1 and B_2 are based on the two simple facts: (1) $Q(u) = \boldsymbol{P}(N(0,1) > u) \leq 1$ for all $-\infty < u < \infty$ and (2) $Q(u) \leq 1/2$ when $u > 0$.

To improve the upper bound of B_2, we propose to further split the subinterval $(\frac{\gamma}{m}, \frac{3\gamma}{2m}]$ into k sub-subintervals:

$$(\frac{\gamma}{m}, \frac{k+2}{k+1}\frac{\gamma}{m}] \text{ and } (\frac{j+2}{j+1}\frac{\gamma}{m}, \frac{j+1}{j}\frac{\gamma}{m}], \text{ for } j = 2, \cdots, k. \tag{2.51}$$

Denote by $B_2^{(1)}$ and $B_2^{(j)}$ the summation as in B_2 in (2.50) except when t is over the first and j-th sub-interval in (2.51), respectively, for $j = 2, \cdots, k$. For the first subinterval in

(2.51), by the simple fact that $\mathbb{Q}(u) \leq \frac{1}{2}$ when $u \geq 0$, we have

$$B_2^{(1)} = \frac{1}{2}\left(\frac{1}{k+1}\frac{\gamma}{m}\right). \tag{2.52}$$

For the j-th subinterval in (2.51) with $j = 2, \cdots, k$, we propose to explore relation (2.37), which provides a much improved bound for $Q(u)$ than the constant $1/2$ when $u > 0$. That is, by (2.37), for $j = 2, \cdots, k$, we have

$$B_2^{(j)} \leq \frac{1}{2} \sum_{\lfloor \frac{j+2}{j+1}\frac{\gamma}{m} \rfloor + 1}^{\lfloor \frac{j+1}{j}\frac{\gamma}{m} \rfloor} \exp\left\{\frac{-N(\gamma - mt)^2}{4mt(Nr^2+1)}\right\} \tag{2.53}$$

Our remaining arguments are similar to those in (2.39), with a minor twist to reflect the change of mean from $\mu_0^*(t)$ to $\mu_1^*(t)$. To be more specific, as in (2.39), it is not difficult to see that

$$\begin{aligned} & \sum_{t=a}^{b} \exp\left\{-c\frac{(\gamma - mt)^2}{mt}\right\} \\ \leq \;& \rho^{-1} \exp\left\{-c\left(\frac{\gamma^2}{mb} - 2\gamma\right)\right\} \exp(-cma) \\ = \;& \rho^{-1}, \end{aligned} \tag{2.54}$$

when $a = \frac{j+2}{j+1}\frac{\gamma}{m}$ and $b = \frac{j+1}{j}\frac{\gamma}{m}$. Combining this with (2.53) yields that

$$B_2^{(j)} \leq \frac{1}{2}\rho^{-1}, \tag{2.55}$$

where the right-hand side upper bound does not depend on γ. By (2.52) and (2.55), we have

$$B_2 \;\leq\; \frac{1}{2}\frac{1}{k+1}\frac{\gamma}{m} + \sum_{j=2}^{k}\frac{1}{2}\rho^{-1}$$

$$= \frac{1}{k+1}\frac{\gamma}{2m} + (k-1)\frac{1}{2}\rho^{-1}. \tag{2.56}$$

Hence, by (2.48), (2.50) and (2.56), the expected sample size of T under H_1 satisfies

$$\begin{aligned}
&\boldsymbol{E}_1(T) \leq B_1 + B_2 + B_3 + B_4 \\
\leq\ & \frac{\gamma}{m} + 1 + \left[\frac{1}{k+1}\frac{\gamma}{2m} + \frac{k-1}{2}\rho^{-1}\right] + \frac{1}{2}\rho^{-1} + \frac{1}{2}\rho^{-1} \\
=\ & \frac{\gamma}{m} + \frac{\gamma}{2(k+1)m} + \frac{k+1}{2}\rho^{-1} + 1, \qquad (2.57)
\end{aligned}$$

for any integer $k \geq 1$. When $k = 1$, this is just the upper bound of $\boldsymbol{E}_1(T)$ in (2.20) derived by [75]. Here we will choose k suitably to drive a better upper bound of $\boldsymbol{E}_1(T)$. In particular, in (2.57), we can choose the integer $k = k^* = \lceil \sqrt{\frac{\gamma\rho}{m}} - 1 \rceil$, and by the similar arguments in (2.46), it is straightforward that (2.24) follows at once from (2.57), which completes the proof of the theorem.

2.6 Conclusion

In this chapter, we investigate the distributed sequential detection problem over a pre-specified network structure. Our focus is on improving the non-asymptotic upper bounds on the error probabilities and expected sample sizes of the CISPRT algorithm proposed by [75]. We derive a tight upper bound through a novel approach to bound the infinite sum of tail probabilities of Gaussian distributions. Our results show that the more the number of sensors or the sparser the network neighborhood connectivity is, the larger the information loss is, i.e., the larger expected sample size is needed to achieve the desired Type I and II error probabilities.

Several future directions can be pursued for distributed sequential detection. First, it

will be interesting to provide more accurate high-order approximations for any network structures, especially when the number of sensors is large. New techniques will likely need to be developed. Second, instead of binary simple hypothesis testing, it will be interesting to develop efficient algorithms when there are nuisance parameters in the alternative hypothesis. Third, here we assume that all local sensors having different distributions simultaneously under the alternative hypothesis H_1, and in some applications, one may want to develop efficient algorithm where only a few unknown subset of local sensors have distributions from H_1. This might be closely related to sparsity detection or false discovery rate in the modern statistics literature. Fourth, in our paper the neighborhood or network structure is pre-specified, and it will be useful to investigate the time-varying network structure. Finally, it will also be interesting to adapt the analysis and the technical tools in this paper to non-linear observations models or non-Gaussian noises.

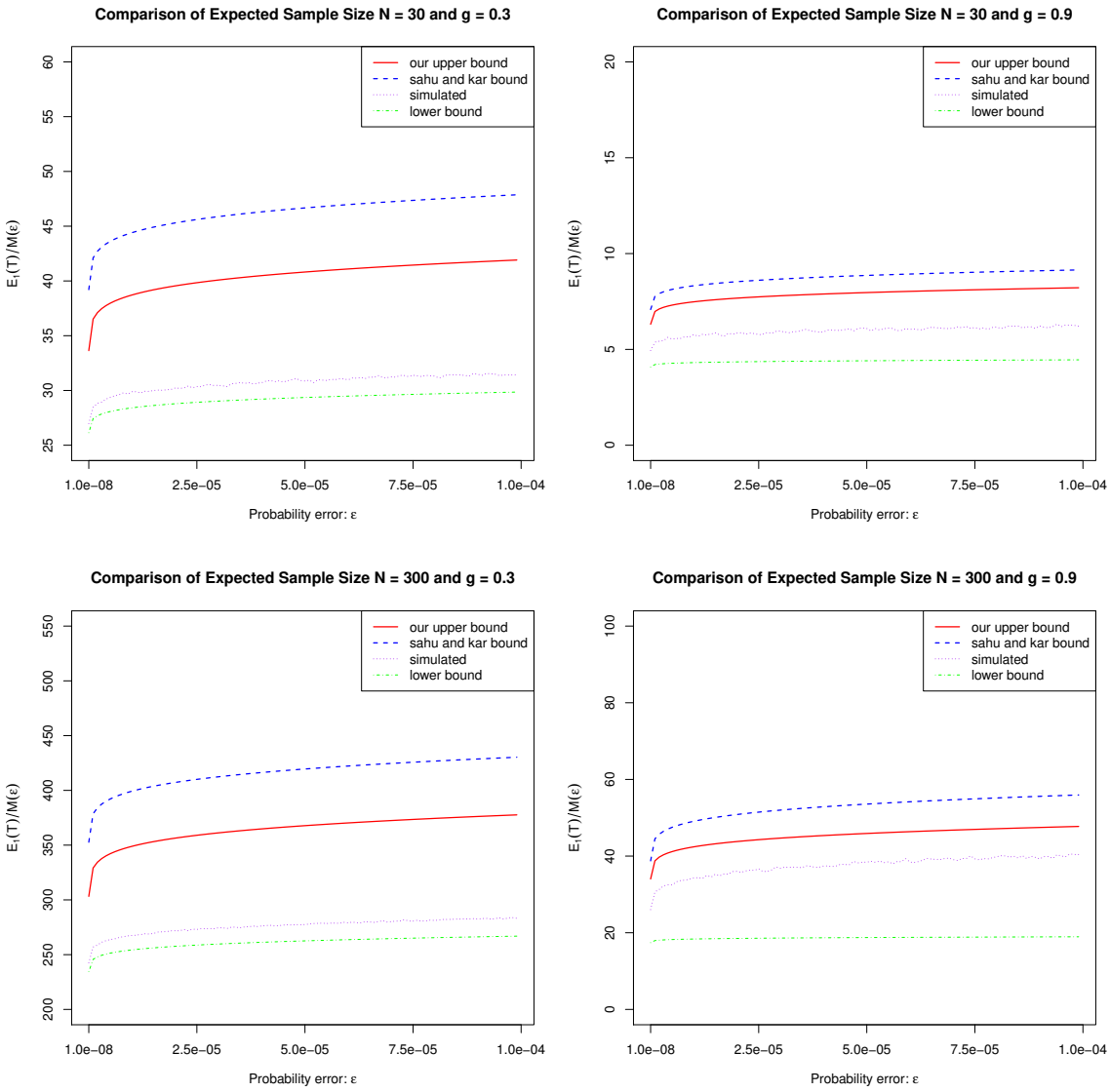

Figure 2.1: A comparison of four different estimates of $\mathbf{E}_1(T_i)/\mathcal{M}(\epsilon)$ of the CISPRT under four different setting of random graph depending on the number N of sensors and the connectivity parameter g. In each plot, four curves represent four different estimates as $\alpha = \beta = \epsilon$ varies, and these four methods ranking from largest to smallest are as follows: (i) The blue dashed line is Sahu and Kar's upper bound in (2.22); (ii) The red solid line is our improved bound in (2.28); (iii) The purple dotted line is the Monte Carlo simulated estimate of $\mathbf{E}_1(T_i)/\mathcal{M}(\epsilon)$; and (iv) The green dotdash line is the lower bound in (2.22). The plots confirms that our bound in (2.28) is attainable when ϵ goes to 0.

CHAPTER 3

EFFECTIVE ASSESSMENT OF TREATMENT EFFECTS IN EARLY ALZHEIMER'S DISEASE THROUGH MIXED-EFFECTS BETA REGRESSION

3.1 Introduction

This research is motivated from an ongoing Phase 3 clinical trial of Aducanumab, the company Biogen's investigational treatment for Alzheimer's disease (AD) that has been shown promising based on the three years of results from the Phase 1b study. This ongoing Phase 3 study started on September 30, 2015, with the estimated primary completion date of February 19, 2020. The primary objective of the study is to evaluate the efficacy of Aducanumab in slowing cognitive and functional impairment as compared with placebo in participants with early AD, and the primary outcome measures are the *changes* in the Clinical Dementia Rating-Sum of Boxes (CDR-SB) score. A crucial statistical question is how to develop an effective statistical test that has a large power to detect the treatment effect of Aducanumab if it is indeed as promising as shown in the Phase 1b study.

This Phase 3 study of Aducanumab poses several statistical challenges. First, the CDR-SB score is not normally distributed, as it is a categorical variable that is highly skewed and takes the values ranging from 0 to 18 with the smallest possible step size of 0.5. Second, the primary outcome measures are not the CDR-SB scores themselves, but the *changes* in the CDR-SB scores from the baselines. One important criterion is the difference on the CDR-SB score between Month 18 and Month 0 after the study, and the current gold standard method is the so-called responder analysis based on the two-sample proportion test, which only uses information at Month 18 and 0. This might lose detection powers due to two reasons: (i) not every subject will have these CDR-SB scores at Month 18, due to various reasons such as missing the appointments or dropping out; and (ii) it does not

take advantage of the longitudinal study design when the CDR-SB scores will be collected multiple times for most subjects (e.g., at Month $0, 6, 12, 18, 24$ and 36 after the enrollment of the study).

The goal of this chapter is to develop an efficient statistical test that is able to effectively assess treatment effects of new drugs in early AD. Our main statistical method is to model the CDR-SB score by the Beta distribution, and to use the mixed-effects Beta regression model to enhance the detection power of the changes in the CDR-SB scores at Month 18 by borrowing information from other time steps such as Months 12 or 24 and by using information from all subjects no matter whether they have missing data at Month 18 or not. Note that much research has been done for longitudinal data analysis [44, 47, 29, 18, 19], and here we propose to apply the mixed-effects Beta regression model proposed by [89, 102] to our context. Beta distribution can fit the observed CDR-SB scores well and can also deal with the heteroskedastic issues that arise when the variance in the CDR-SB scores decreases over time (to approach 0) as the scores approach the upper or lower boundaries of the scale. See [17] for more arguments of how beta regression models was proposed, and [5] for the beta regression with applications in the medical research. More recently, beta regression was used for fitting the longitudinal categorical response, see [34, 98, 77].

The remainder of this chapter is as follows. In Section 3.2, we formulate the problem of accessing the treatment effects of new drugs in early AD and the gold standard methods are introduced and discussed. In Section 3.3, we propose the mixed-effects Beta regression model for the CDR-SB. Real data set for the placebo group is presented in Section 3.4. In Section 3.5, we present the simulation studies and the comparison results between the gold standard method and our proposed method.

3.2 Problem Formulation and Gold Standard Methods

In this section, we present the background of the CDR-SB score for the AD patients, as well as the existing golden standard method, the responder analysis, that assesses treatment effects of new drugs in early AD.

The CDR-SB is a useful criterion to measure the progression of AD, see [11, 8, 94, 35]. It has six domains or components: three in cognitive domains (Memory, Orientation, Judgment and Problem Solving) and three in functional domains (Community Affairs, Home and Hobbies, Personal Care). The first five domains are scored on a five-point ordinal scale $(0, 0.5, 1, 2, 3)$, and the last domain is scored on a four-point scale $(0, 1, 2, 3)$. The CDR-SB, the summed score, ranging from 0 to 18, is the total estimate of dementia severity.

In the Phase 3 clinical trial of Aducanumab, there are two groups of subjects: one is the placebo (PBO) group, and the other is the treatment (TRT) group. The CDR-SB scores per subject are observed over time, and the primary outcome measures are the *changes* in the CDR-SB score at Month 18 from the baseline:

$$d = \text{CDR-SB}_{\text{Month18}} - \text{CDR-SB}_{\text{Month0}}. \tag{3.1}$$

Denote by μ_{PBO} and μ_{TRT} the expected CDR-SB change value d in (3.1) for the PBO and TRT groups, respectively. Since it is expected that Aducanumab will slow down the AD progress, the Phase 3 clinical trial of Aducanumab essentially tests the hypothesis

$$H_0 : \mu_{\text{PBO}} - \mu_{\text{TRT}} = 0 \quad \text{versus} \quad H_1 : \mu_{\text{PBO}} - \mu_{\text{TRT}} = \Delta, \tag{3.2}$$

where $\Delta > 0$ is a pre-specified constant, e.g., the data in the Phase 1b study of Aducanumab suggests $\Delta = 0.5$.

In the context of assessing treatment effects in early AD, the gold standard method for testing the hypothesis in (3.2) is the so-called responder analysis which is based on the two-

sample proportion tests. It is assumed that we observe M_1 CDR-SB change value d in (3.1) for the PBO group, say, $d_{1,i}$ for $i = 1, \cdots, M_1$, and observe M_2 CDR-SB change value d in (3.1) for the TRT group, say, $d_{2,j}$ for $j = 1, \cdots, M_2$. While we can assume that the $d_{1,i}$'s are independent and identically distributed (i.i.d.) with mean μ_{PBO} and the $d_{2,j}$'s are i.i.d. with mean μ_{TRT}, the classical two-sample t-test or nonparametric Wilcoxon-Mann-Whitney rank-sum test are not suitable in our context due to the inappropriate normal distribution assumption, many tie values, or the lower power. The main idea of the gold standard responder analysis approach is to dichotomize subjects into "responders" and "non-responders" depending on whether the observed CDR-SB change values d's exceed a suitable threshold value or not, see [24, 62, 63, 81, 87] for the motivations and applications.

To be more specific, the responder analysis chooses a suitable cutoff C for the CDR-SB change value d in (3.1) to indicate whether a subject's AD status progresses significantly worse or not, and define

$$\hat{p}_1 = \frac{1}{M_1}\sum_{i=1}^{M_1} \boldsymbol{I}(d_{1,i} > C) \quad \text{and} \quad \hat{p}_2 = \frac{1}{M_2}\sum_{j=1}^{M_2} \boldsymbol{I}(d_{2,j} > C). \tag{3.3}$$

The problem in (3.2) can then be re-formulated as testing the one-sided hypothesis

$$H_0' : p_1 = p_2 \quad \text{versus} \quad H_1' : p_1 > p_2, \tag{3.4}$$

and the corresponding test statistic is given by

$$z_{\text{obs}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{M_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{M_2}}}, \tag{3.5}$$

which asymptotically follows the standard normal distribution. In particular, at the significant level $\alpha = 2.5\%$, the responder analysis will claim there is a significant difference between PBO and TRT groups if and only if $z_{obs} > z_\alpha = 1.96$.

There are several issues to adopt the gold standard responder analysis in practice. First,

besides the lower power due to dichotomize the observations, the main challenge is how to choose the cutoff value C in (3.3) suitably so as to maximize the detection power under H_1 in (3.2). Had the data been the normally distributed, it would be straightforward to show that $C = (\mu_{\text{PBO}} + \mu_{\text{TRT}})/2 = \mu_{\text{PBO}} + \Delta/2$ under (3.2). However, since the CDR-SB data is not necessarily normally distributed, it might be nontrivial to understand the impact of the cutoff value C on the detection power, which will be investigated in our simulation studies. Second, the power of responder analysis will significantly be affected if there are missing observed CDR-SB values at Month 18. Third, it does not take advantage of the multiple observed CDR-SB values per subject over time. Our objective is to develop a more powerful test than the gold standard responder analysis.

3.3 Our Models and Methods

In the context of Phase 3 study of Aducanumab, suppose there are M subjects in the longitudinal dataset, where $i = 1, 2, \ldots, M$. For the i-th subject, a CDR-SB score $Y^*_{i,j}$ is observed at time $t_{i,j}$ (in years) for $j = 1, 2, \ldots, n_i$. E.g., $t_{i,j} = 1.5$ corresponds to the observation at Month 18 in the study. The subjects are divided into two groups, PBO and TRT groups, which can be represented as the binary indicator variable $x_i \in \{0, 1\}$, with $x_i = 1$ being from the TRT group. Thus our observed longitudinal data can be written as (Y^*_{ij}, t_{ij}, x_i) for $i = 1, \cdots, M$ and $j = 1, \cdots, n_i$.

We propose a mixed-effects Beta regression model to assess treatment effects in early AD. For that purpose, note that the CDR-SB scores take the values in $[0, 18]$ but the beta distribution is defined in the open range of $(0, 1)$. Thus we propose to first transform the response CDR-SB scores to

$$Y_{i,j} = \frac{(1 - r)Y^*_{i,j}}{18} + \frac{r}{2}, \tag{3.6}$$

where r is a pre-sepcified small positive value that allows the boundary CDR-SB scores to

be inside the interval $(0, 1)$ after the transformation. In our case studies and simulations, we choose $r = 0.01$ for simplicity.

Next, conditional on covariates (t_{ij}, x_i), we assume that the responses $Y_{i,j}$ in (3.6) follows a beta distribution $Beta(\alpha, \beta)$ with $\alpha = \mu_{i,j}\phi$ and $\beta = (1 - \mu_{i,j})\phi$, which has

$$\mathbf{E}(Y_{i,j}) = \frac{\alpha}{\alpha + \beta} = \mu_{i,j} \text{ and } Var(Y_{i,j}) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} = \frac{\mu_{i,j}(1 - \mu_{i,j})}{1 + \phi}. \quad (3.7)$$

Then the logit transformation of the mean μ_{ij} can be modeled as linear mixed effects of the covariates (t_{ij}, x_i). To be more specific, our proposed mixed-effects Beta regression model is as follows.

$$\begin{aligned} &Y_{ij}|(t_{ij}, x_i) \sim \text{Beta}(\mu_{i,j}\phi, (1 - \mu_{i,j})\phi), \text{ with} \\ &\log\left(\frac{\mu_{i,j}}{1 - \mu_{i,j}}\right) = \beta_0 + \beta_1 t_{i,j} + \beta_2 x_i + \beta_3 t_{i,j} x_i + b_{0,i} + t_{i,j} b_{1,i}, \end{aligned} \quad (3.8)$$

where the β_k's are fixed effects, and $(b_{0,i}, b_{1,i})$ are two random effects in the sense that $(b_{0,i}, b_{1,i})$ are bivariate normal distributed with $N((0, 0), G)$ for some 2×2 covariance matrix G.

It is useful to interpret our proposed model (3.8) back to the original hypothesis context in (3.2) on the CDR-SB change score. Note that $t_{ij} = 1.5$ and 0 corresponds to the time of Month 18 and 0, respectively. Letting $x_i = 0$ or 1, under our proposed model (3.8), we have

$$\begin{aligned} \mu_{\text{PBO}} &= \frac{1}{1 + \exp(-\beta_0 - 1.5\beta_1)} - \frac{1}{1 + \exp(-\beta_0)} \quad \text{and} \\ \mu_{\text{TRT}} &= \frac{1}{1 + \exp(-\beta_0 - \beta_2 - 1.5(\beta_1 + \beta_3))} - \frac{1}{1 + \exp(-\beta_0 - \beta_2)}. \end{aligned} \quad (3.9)$$

Moreover, note that in the context of Phase 3 study of Aducanumab, the patients are randomized into the PBO and TRT groups, and thus the CDR-SB scores at Month 0 are assumed to be the same for the PBO and TRT groups. In other words, theoretically we should

expect $\beta_2 = 0$. In the special case of $\beta_2 = 0$, we have

$$\mu_{\text{PBO}} - \mu_{\text{TRT}} = \frac{1}{1+\exp(-\beta_0 - 1.5\beta_1)} - \frac{1}{1+\exp(-\beta_0 - 1.5(\beta_1+\beta_3))}, \tag{3.10}$$

and thus the original hypothesis context in (3.2) reduces to test

$$H_0^* : \beta_3 = 0 \quad \text{versus} \quad H_1^* : \beta_3 < 0 \tag{3.11}$$

In our simulation studies below, we will assume $\beta_2 = 0$ to generate simulated data. However, when we fit the real or simulated data, we keep the term β_2 in (3.8), and then conduct the hypothesis (3.11) on β_3 to see whether there is a significant treatment effect or not. In particular, by the asymptotic properties of the maximum likelihood estimators (MLE), one would reject $H_0^* : \beta_3 = 0$ at the significant level α (e.g., $\alpha = 2.5\%$) if

$$Z_{\text{obs}} < -z_{\alpha/2}, \text{ where the test statistic } Z_{\text{obs}} = \hat{\beta}_3 / se(\hat{\beta}_3),$$

where $\hat{\beta}_3$ and $se(\hat{\beta}_3)$ are the MLE and the corresponding standard error of the parameter β_3, and $z_{\alpha/2}$ is the cutoff value z such that $\mathbf{P}(N(0,1) > z) = \alpha/2$.

We should mention that it is nontrivial to fit our proposed mixed-effects Beta regression model in (3.8). In R, the traditional *glm()* family does not include the beta distribution, although the beta regression might be fitted by a separate package, such as *betareg* by [12]. Unfortunately, for the longitudinal data, currently there is no available traditional R package to fit mixed-effects Beta regression model. For the SAS software, we found that it can fit our model through the *GLIMMIX* procedure for Generalized Linear Mixed models, but the algorithms sometimes fail to converge, partly because of the limitations of the optimization algorithm. In addition, the covariance matrix $\hat{G}$ of the random effects $(b_{0,t}, b_{1,t})$ can only be the diagonal matrix for Beta distribution in the SAS.

In our numerical simulations below, we also implement Bayesian-type estimates by

using the programming language, *Stan*, which allows us to fit arbitrarily complex models through the Bayesian procedure. We implemented the *Stan* program in R via the *RStan*, see [23, 82] for more descriptions and information. When fitting our proposed models via Bayesian methods, we need to specify the prior distributions of the unknown parameters $(\beta_0, \beta_1, \phi, G)$, and the prior distributions are chosen as follows: both β_0 and β_1 follow $N(0, 100)$; ϕ follows $Gamma(0.01, 0.01)$; and the covariance matrix G follows either inverse wishart $W^{-1}(2, \mathbf{I})$ or two diagonal elements of G follow $Uniform(0, 10)$. It turns out that our Bayesian estimates always converge. Moreover, they are almost identical to the MLEs in the *SAS* when the latter does converge.

3.4 Real Data Set for the PBO Group

We do not have access to the real data set from the Phase 3 study of Aducanumab which is ongoing. Here we used public datasets which can be thought of as the placebo group. This will allow us to fit our proposed mixed-effects Beta regression model (3.8) to the placebo group, and then use relation (3.10) to simulate data for the TRT group to compare detection powers of different methods.

The real data set used for our statistical analysis was obtained on June 1, 2017 from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database funded by the National Institute of Health (NIH), see the link (`http://adni.loni.usc.edu`). Also see [93] for the background, goals, and structure of the ADNI database. We focus on the CDR-SR relevant data in the ENGAGE/EMERGE study of the ADNI database, which has 12742 observations from 1737 subjects over up to 10 years.

Our first step is to pre-process the raw data to select a subgroup of subjects that mimic the populations in the Phase 3 study of Aducanumab. This subgroup can be thought of as a representative of the placebo (PBO) group, and will allow us to capture the natural disease progression for the early stage of AD patients. The specific eligibility criteria of the chosen subjects are as follows.

1. Diagnostic baseline is Late MCI (LMCI) or AD.

2. Aged 50 to 85 years old, inclusive, at the time of informed consent.

3. Must have at least 6 years of education or work experience to exclude mental deficits other than Mild Cognitive Impairment (MCI) or mild AD.

4. Must have a positive amyloid Positron Emission Tomography (PET) scan. This is defined as meeting at least one criterion hierarchically as below:

   (a) Amyloid PET. If AV-45 is greater or equal to 1.13, and if AV-45 is missing, then consider PIB greater or equal to 1.47.

   (b) CerebroSpinal Fluid (CSF). If PET is missing, then log ratio (logarithm of p-Tau/β-amyloid) > -3.6 (equivalent to p-Tau/β-amyloid > 0.028).

   The cut-off values can be determined by computing one standard deviation below the average of normal subjects.

5. Must meet all of the following clinical criteria for MCI due to AD or mild AD according to NIA-AA criteria [1, 56], and must have:

   (a) A CDR-Global Score of 0.5.

   (b) A Repeatable Battery for the Assessment of Neuropsychological Status (RBANS) equivalent cognition test score. Since RBANS were not available in ADNI, an equivalent test is used, see more in [30]. ADNI used Logical Memory, Delayed Recall at Baseline to define MCI, but according to literature, this criteria is perhaps not as stringent as RBANS. Therefore, additional cognition criteria is used, based on Auditory Verbal Learning Test (AVLT) Version B - Delayed, 30 Minute Delay Total at Baseline (variable named as AVLTCTBL) below < 3.67 which is 1 standard deviation below the average of normal people.

(c) A Mini-Mental State Examination (MMSE) score between 24 and 30 (inclusive).

Note that Criterion (1) allows us to focus more on the progression of early AD, and other criteria make sure that our selected subjects mimic the placebo (PBO) groups in the Phase 3 study of Aducanumab.

The final selected dataset consists of 1206 observations collected from 237 subjects over up to 10 years. Figure 3.1 shows the distributions of CDR-SB for all months and for Month 0. For the histograms, it is clear that the distribution of the CDR-SB scores is highly skewed to the left, and thus the normality assumption indeed might be inappropriate.

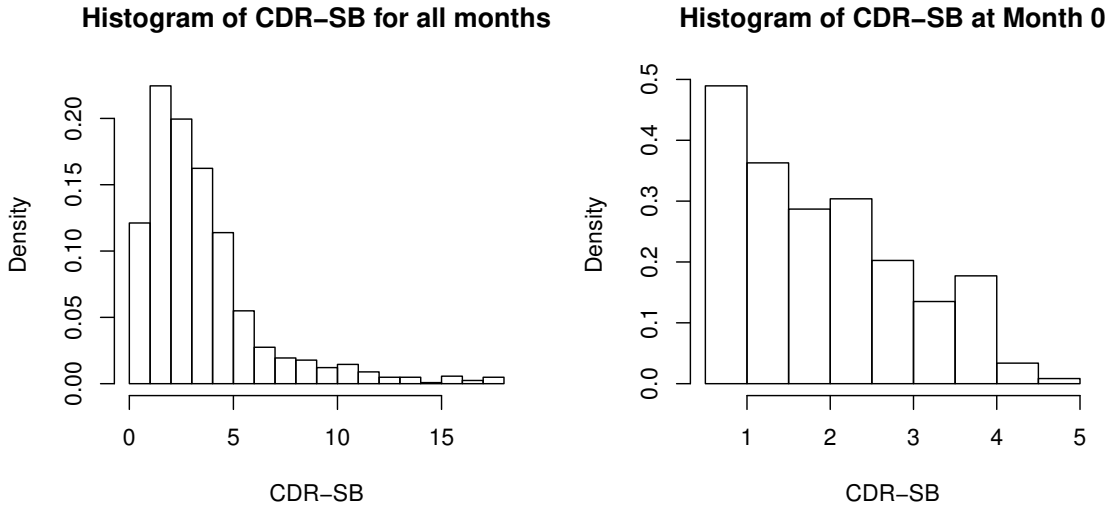

Figure 3.1: Histograms of all CDR-SB scores and the CDR-SB scores at Month 0, respectively.

Moreover, Figure 3.2 illustrates the longitudinal progression of the CDR-SB scores for the PBO group over time, and Table 3.1 includes some summary statistics of the CDR-SB scores in first 36 months for all selected 237 subjects. From the plot and table, the mean of the CDR-SB scores increases almost 0.89 in the first year, and increases almost 1.23 in the second year, which shows that the disease will progress faster as time goes by. Moreover, the effective sample sizes are decreasing from $N = 237$ at Month 0 to $N = 124$ at Month

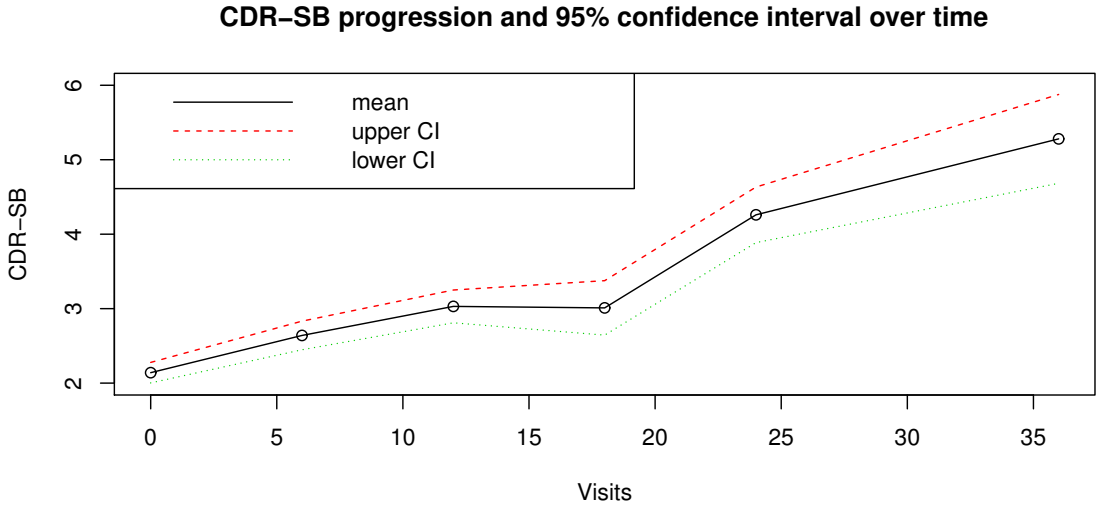

Figure 3.2: CDR-SB progression and its 95% confidence interval over 3 years.

Table 3.1: Observed CDR-SB in first 36 months for selected group.

	CDR-SB		
Month	N	mean	sd
Month 0	237	2.14	1.08
Month 6	231	2.64	1.49
Month 12	216	3.03	1.65
Month 18	86	3.01	1.73
Month 24	178	4.26	2.54
Month 36	124	5.28	3.39

Table 3.2: The estimated parameters in (3.8) using the ADNI dataset.

Estimated parameter	Fitted value
$\hat{\beta}_0$	-2.0646
$\hat{\beta}_1$	0.4244
$\hat{\phi}$	80.5312
$\hat{G}$	$\begin{pmatrix} 0.2549 & 0 \\ 0 & 0.08338 \end{pmatrix}$

36. Note that many subjects do not have the CDR-SB value at Month 18, although we are interested in the treatment effect of Aducanumab at Month 18.

Finally, we fit our proposed mixed-effects Beta regression model in (3.8) to these $N = 237$ patients from the PBO group with $x_i = 0$, and the estimated parameters, $\hat{\beta}_0, \hat{\beta}_1, \hat{\phi}$

Table 3.3: The comparison of the fit statistics between the mixed effects Beta regression model and LMM (smaller is better).

Fit statistics	mixed-effects Beta regression	LMM
MSE: $\sum_{i=1}^{M}\sum_{j=1}^{n_i}(Y_{i,j}-\hat{Y}_{i,j})^2$	797.44	973.05
AIC	-2662.99	-3099.83
AICC	-2662.94	-3099.78
BIC	-2645.77	-3082.62
CAIC	-2640.77	-3077.62
HQIC	-2656.04	-3092.89

and the covariance matrix $\hat{G}$ of $(b_{0,i}, b_{1,i})$ are summarized in Table 3.2. In addition, the convergence plots of $\hat{\beta}_0, \hat{\beta}_1$ with 1000 Monte Carlo iterations are presented in Figure 3.3.

As a comparison, the standard linear mixed-effects model (LMM) is also fit to the same PBO group. Figure 3.4 shows two longitudinal examples over 120 months, which indicates that the mixed-effects Beta regression model fits the actual data better and can capture the possible curvature of the response. Figure 3.5 presents the mean of actual and fitted CDR-SB through the LMM model and our proposed mixed-effects Beta regression model. Moreover, Table 3.3 reports the fitted statistics between the LMM and our proposed mixed-effects Beta regression model. These plots and tables suggest that our mixed-effects Beta regression model fits the observed data set well.

3.5 Simulation Studies

In this section we report the results of our simulation studies to show the usefulness of our proposed mixed-effect Beta regression model. For better understanding, we split it into two subsections: (1) generative models for both PBO and TRT groups; and (2) numerical simulation results show that our proposed mixed-effect Beta regression model is more powerful than the gold standard responder analysis method.

3.5.1 Generative models

Our simulations are based on the real data in Section 3.4. In our Monte Carlo simulation, we generate a simulated data set of $M = 400$ subjects: half for PBO group and half for TRT group. In the ideal full data context, each subject has five visits at month 0, 6, 12, 18, 24, respectively, which corresponds to $t_{i,j} \in \{0, 0.5, 1, 1.5, 2\}$ in years. For the PBO group, we simulate the observed CDR-SB values per subject from our proposed mixed-effects Beta regression model (3.8) with the parameters $(\hat{\beta}_0, \hat{\beta}_1, \hat{\phi}, \hat{G})$ in Table 3.2. For the TRT group, we simulate the observed CDR-SB values per subject from the model (3.8) with $x_i = 1$, $\beta_2 = 0$ and β_3 is given by (3.10) when $\mu_{\text{PBO}} - \mu_{\text{TRT}} = \Delta$ under H_1 in (3.2).

In our simulation, we consider two values: $\Delta = 0.5$ and 0.7, and for each given Δ value, we run 1000 Monte Carlo runs to get the statistical powers. In addition, note that when $\Delta = 0$, the generative models for both PBO and TRT groups are the same, and thus this will allow us to validate that our test is a significance level $\alpha = 2.5\%$ test.

Moreover, the missing data case is used to compare with the full data case, since we would like to see if the number of data points at Month 18 decreases, how it will affect the statistical power. Motivated by the real data set in Section 3.4, we introduce 20% of missing subjects to each simulated dataset. It is assumed that for both PBO and TRT group, there are no missing observations at Month 0, $20 \times \frac{1}{3}$ percent of missing observations at Month 6, $20 \times \frac{1}{3}$ percent of missing observations added at Month 12, and another $20 \times \frac{1}{3}$ percent of missing observations added at Month 18. Hence, at Month 18, 20% of missing observations exist in the simulated dataset. We should mention that our simulation setup reflects the real data, since more and more people often drop off the clinical trials along with the study.

Table 3.4: Percentages of events and statistical powers for responder analysis are reported under both no missing data and 20% missing cases.

No missing values in PBO and TRT groups						
Δ	Gold Standard Method					Our Proposed Method
	Different cutoff C in (3.3)					
	0	0.5	1	1.5	2	
0.5	0.53	0.66	**0.72**	**0.70**	0.68	0.88
0.7	0.84	0.92	**0.94**	**0.94**	0.93	1.00
20% missing in both PBO and TRT groups						
0.5	0.42	**0.54**	**0.56**	0.51	0.44	0.82
0.7	0.74	**0.81**	**0.84**	0.79	0.71	0.99

3.5.2 Numerical Results

Table 3.4 compares the powers of our proposed methods with the gold standard methods at the significance level $\alpha = 2.5\%$ based on the 1000 Monte Carlo runs. For the complete full data set with $M = 200$ subjects per group, the gold standard responder analysis methods have good powers, and our proposed mixed-effects Beta regression methods are moderately better. When there is missing data, the powers of our proposed mixed-effects Beta regression method are much larger than the gold standard responder analysis methods. For instance, when $\Delta = 0.5$, the highest power for the gold standard method is 0.56, whereas the power for our proposed method is 0.82. Moreover, the optimal choice of the cutoff C in the gold standard method seems to depend on the target difference Δ in (3.2) as well as the proportion of missing data, which indicates that the use of the gold standard method can have much lower power in practice if the cutoff value C is not suitably chosen.

3.6 Discussion

In this chapter we developed mixed-effect Beta regression models to fit observed CDR-SB values for early AD patients. Our models and methods fit a real data set well for the PBO group as compared to the standard LMM models, and can increase powers for detecting treatment effect of new drugs as compared to the gold standard responder analysis methods.

There are several interesting problems that deserve further research. While our proposed mixed-effect Beta regression models seem reasonable for the specific subgroups of early AD patients over the duration of two years, it will be interesting to see whether our proposed models can be extended to clinical studies or observational studies with longer durations or more general populations. There are some potentials for such extensions, but some twists are unavoidable. For instance, the linear trend will be too restrict for the longer period, and we might consider piecewise linear or other nonlinear kernel smooth for the temporal trend of CDR-SB values per subject. Also it will be interesting to include more explanatory variables in our models so as to better investigate at the individual level. This will likely involve balancing the tradeoff between individual level efficiency and global level model robustness. Hopefully our research opens some new directions to model and analyze CDR-SB values for early AD patients.

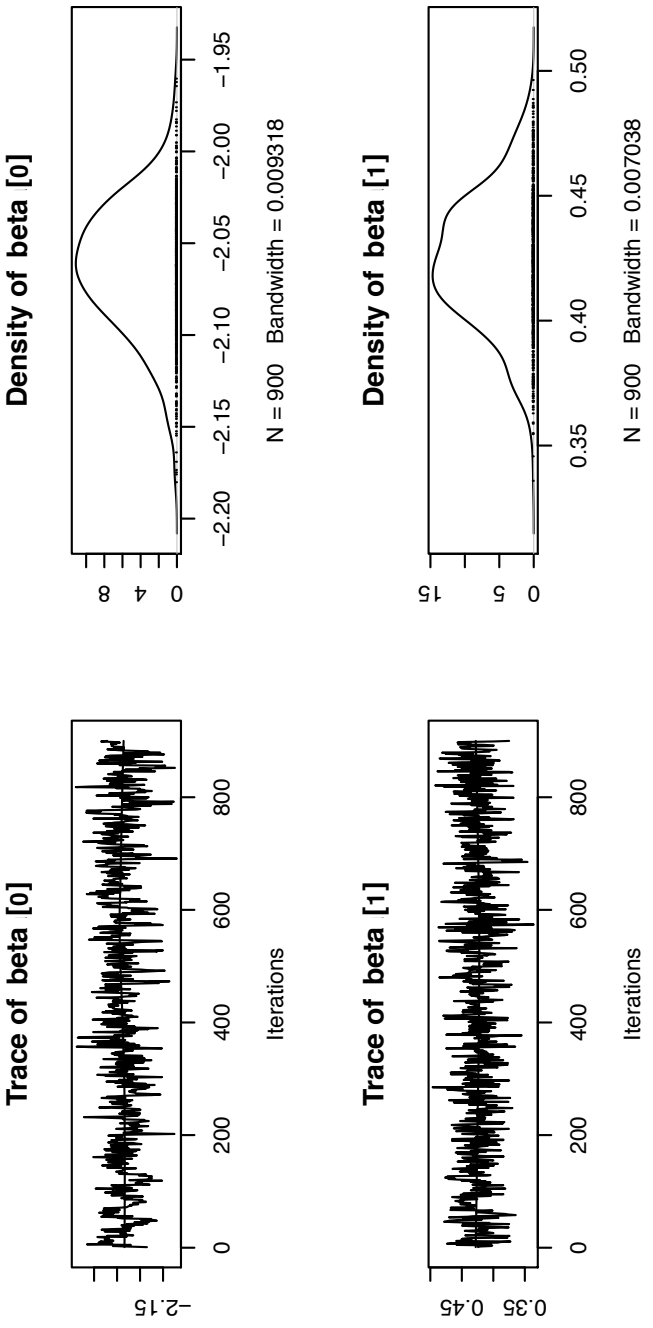

Figure 3.3: Convergence plots of the intercept β_0 and the slope β_1 using the Bayesian method based on 1000 Monte Carlo runs.

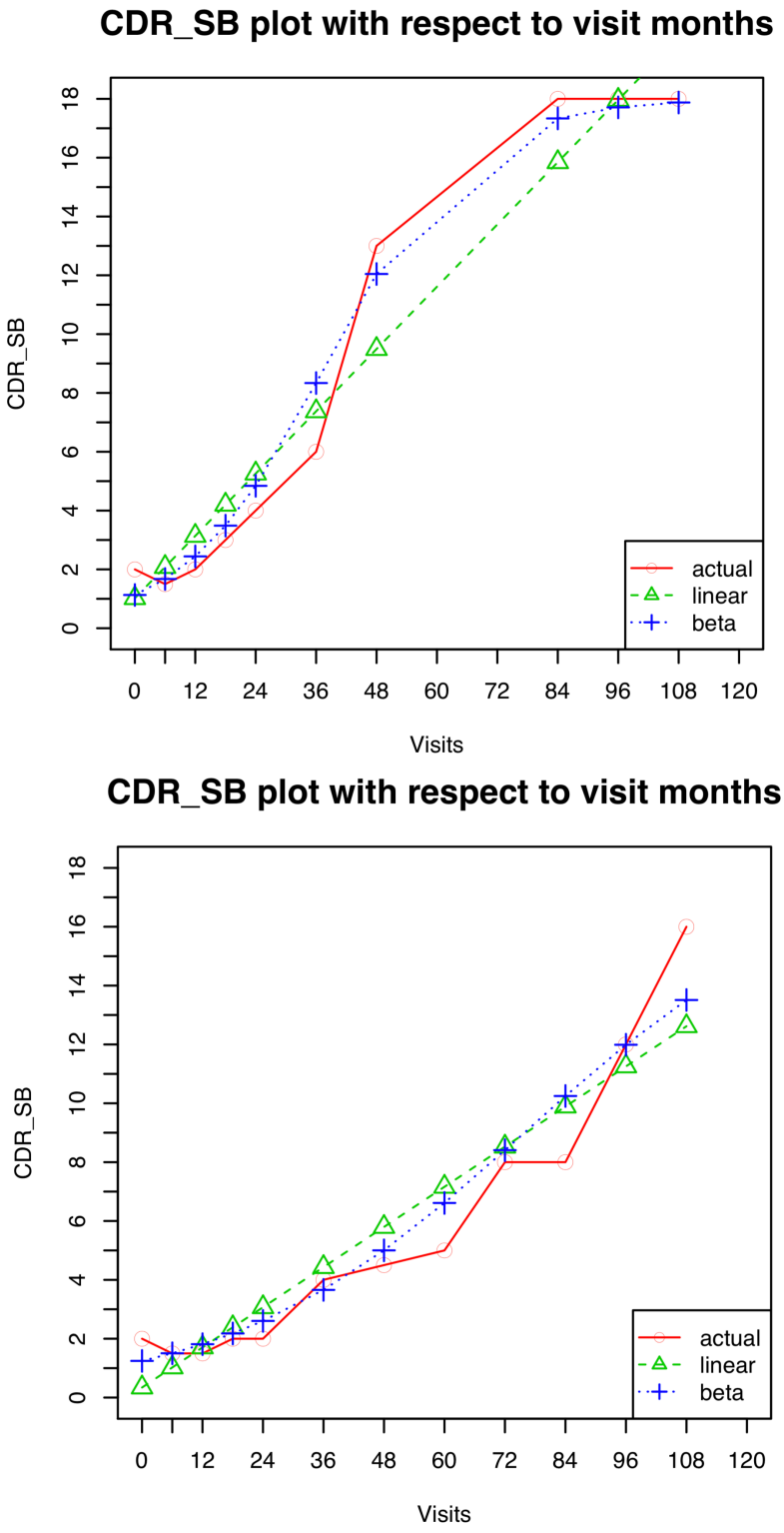

Figure 3.4: Selected longitudinal examples to illustrate that our proposed mixed-effects Beta regression model fits the data well as compared to the standard linear mixed-effects model (LMM).

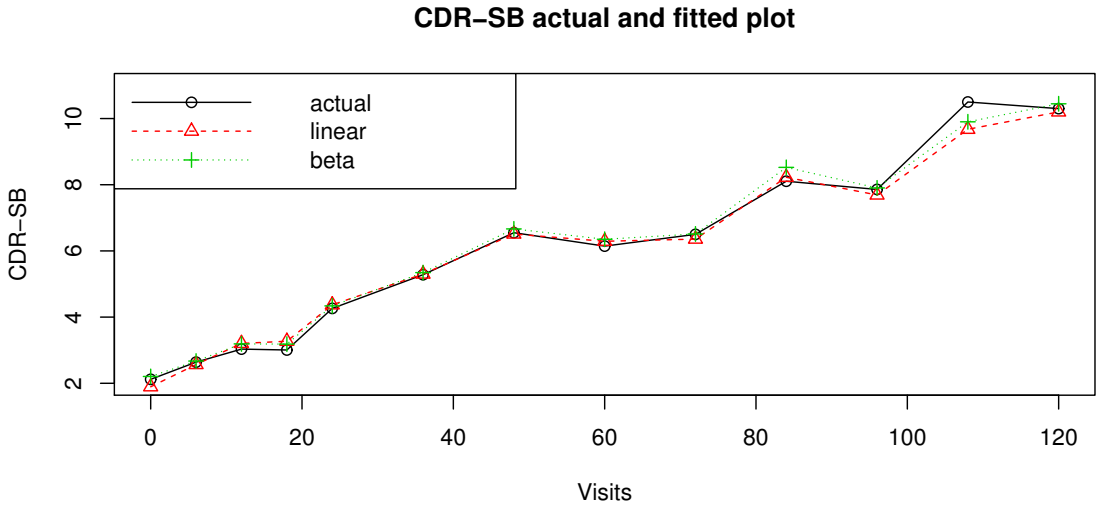

Figure 3.5: CDR-SB actual mean plot and comparison with fitted value through LMM and mixed-effect Beta regression model.

CHAPTER 4

ROBUST ESTIMATION UNDER EXPONENTIAL LOSS FUNCTION

4.1 Introduction

Data integration from different sources has a wide range of real-world applications including sensor networks, image and video processing, pattern classifiers, automatic detection [40, 20], as it allows one to make a more accurate decision at the global level by combining different local knowledge together. One of the main challenges to develop efficient data integration techniques is the existence of outliers and spurious data, which are often caused either by the natural environment noises or by the malicious sensors in adversarial machine learning. In order for the global decision to be efficient and resistant to the potential outliers and spurious data, it is crucial for each local data source to make a robust local inference.

This chapter investigates a novel simple but useful robust data integration technique in the specific context of parameter estimation and linear regression. Note that we do not aim to develop general robust data integration techniques, but want to illustrate the usefulness of classical robust statistics pioneered by Huber [31, 33, 32] in some modern applications, also see [9, 51] for more discussions on the robust methods and their properties in high-dimensional settings.

Our main idea is to propose an exponential loss function to bound the effects of outliers. Note that under the traditional L^2 or Huber's loss function, the effects of outliers can be unbounded, and thus the corresponding estimators might become inaccurate as compared to the case where we know which data points are outliers/contaminated and which are not. Our approach is motivated by the maximum Lq-likelihood estimator [16, 70], which can be thought of robustifying the maximum likelihood estimator in the presence of outliers, as the exponential loss function can be better to mitigate the effect of outliers if well-designed.

Our main contribution is to develop an asymptotic theory in a new asymptotic regime when the outlier means go to ∞ in a suitable rate as the proportion of outliers goes to 0.

The remainder of this chapter is organized as follows. In Section 4.2, we present the problem formulation and the motivation of the mixture model. Section 4.3 proposed our L^α estimator for the mixture model and discussed the properties of our proposed estimator in both traditional and modern asymptotic regimes. In Section 4.4, our proposed estimator is investigated in the regression models and simulation studies are conducted to compare the performance of our proposed method with the traditional MLE and Huber's estimator. Some concluding remarks and further directions are included in Section 4.5.

4.2 Problem Formulation

Assume we observe a set of N independent observations, say, $\{Y_1, Y_2, \ldots, Y_N\}$, which includes a proportion ϵ of outliers. Sine we do not know which observations are outliers, we might assume that the observations are i.i.d. from a two-component mixture model with probability density function (pdf) of $(1-\epsilon)f_{\mu_0}(\cdot)+\epsilon g(\cdot)$, where f_{μ_0} is the parametric pdf we are interested in for the most observations, and g is the pdf of the outliers or contaminated observations. Our main objective is to efficiently estimate the parameter μ_0 in f_{μ_0} based on the contaminated data $Y_1, \ldots, Y_N$.

Note that extensive research has been done in the classical robust statistics literature when $g(\cdot)$ is a heavy tail distribution centered around μ_0, and here we focus on a different case where the mean of $g(\cdot)$ is deviated from μ_0. One natural idea is to first identify which observations are from the interested parameter model f_{μ_0} and which observations are from the contamination model g, and then only use the observations from f_{μ_0} to estimate μ_0. This approach might work for low dimension, but unfortunately it can be very challenging to cluster the observed data correctly into two clusters in the high dimensional setting, since the data becomes much sparser in the high dimensional setting.

Here we propose to follow the traditional M-estimator to suppress the outlier effect,

and one main difference is to choose a loss function that has a bounded loss no matter how extreme the outliers are. The rationale is that we do not aim to derive an unbiased estimator of μ_0 for finite-sample settings, but want to develop an estimator that can be asymptotically unbiased.

In the remaining of this section, let us review the traditional M-estimator and the mixture model. At the high-level, the M-estimator is the solution of the following optimization problem:

$$\hat{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} \rho(||Y_i - \mu||), \tag{4.1}$$

where $\rho(u)$ is some reasonable loss function that is able to control the outlier effects. Note that the maximum likelihood estimator can be written in the form in (4.1), and thus can be thought of as a special case of M-estimator if one prefers.

Note that one of the most widely used loss function is the Huber loss function:

$$\rho(u) = \begin{cases} \frac{1}{2}u^2 & \text{if } |u| \leq \lambda, \\ \lambda(|u| - \frac{1}{2}\lambda) & \text{if } |u| > \lambda. \end{cases} \tag{4.2}$$

From the modelling viewpoint, Huber's M-estimator is the Lasso estimator of μ_1 under the following model

$$Y_i = \mu_1 + \Delta_i + \epsilon_i, \tag{4.3}$$

where $\epsilon_i \sim N(0, \sigma^2)$ and $(\Delta_1, \ldots, \Delta_n)$ is a sparse vector with nonzero elements representing the outlier data. The Lasso estimator of μ_1 is given by

$$\hat{\beta}_{Lasso} = \arg\min_{\beta, \Delta_i} \left[\sum_{i=1}^{n} (Y_i - \mu_1 - \Delta_i)^2 + \lambda|\Delta_i| \right],$$

which reduces to the M-estimator in (4.1) under the Huber loss function (4.2).

In our context, we consider a two-component mixture model. Let us use multivariate normal distributions as an illustration, though the idea can be easily generalized to any other distributions. Assume that f_{μ_0} is the pdf of the multivariate-normal distribution with mean vector μ_1 and covariance matrix Σ_1, and g is the pdf of the multivariate-normal distribution with mean vector $\mu_2 = \mu_1 + \Delta$ and variance Σ_2. Alternatively, it is useful to introduce an unobservable indicator variable, δ_i, to indicate whether the i-th observation Y_i is from the interested component f_{μ_0} or not. The mixture model can then be rewritten as follows: $P(\delta_i = 0) = 1 - \epsilon$ and $P(\delta_i = 1) = \epsilon$, and the observation Y_i satisfies the model

$$Y_i = \mu_1 + \delta_i \Delta + \delta_i (\Sigma_2^{1/2} \Sigma_1^{-1/2}) z_i, \tag{4.4}$$

where $z_i \sim N(0, \Sigma_1)$. Here we want to estimate μ_1 under the assumption that Σ_1 and Σ_2 are pre-specified, and Δ is unknown.

Had we observed the indicator variables δ_i's, the "complete" likelihood function would become

$$\begin{aligned} L(\mu) &= -\sum_{i=1}^{n} \frac{1}{2} \left((Y_i - \mu)^T \Sigma_1^{-1} (Y_i - \mu) + p \log(2\pi) + \log(|\Sigma_1|) \right) \delta_i \\ &\quad -\sum_{i=1}^{n} \frac{1}{2} \left((Y_i - \mu - \Delta)^T \Sigma_2^{-1} (Y_i - \mu - \Delta) + p \log(2\pi) + \log(|\Sigma_2|) \right) (1 - \delta_i). \end{aligned} \tag{4.5}$$

and thus the corresponding "complete MLE" would become

$$\begin{aligned} \hat{\mu} &= \arg\min_{\mu, \Delta} \sum_{i=1}^{n} \left[(Y_i - \mu)^T \Sigma_1^{-1} (Y_i - \mu) \delta_i \right. \\ &\quad \left. + (Y_i - \mu - \Delta)^T \Sigma_2^{-1} (Y_i - \mu - \Delta)(1 - \delta_i) \right] \\ &= \arg\min_{\mu} \sum_{i=1}^{n} [(Y_i - \mu)^T \Sigma_1^{-1} (Y_i - \mu) 1(\delta_i = 1)]. \end{aligned} \tag{4.6}$$

This essentially said that we should only use the observations from the f_{μ_0} component to make inference μ_0, and it makes senses since we do not know the impacts of the outliers as

$\Delta = \mu_2 - \mu_1$ is unknown.

When Σ_1 is identical matrix and the outliers have large mean deviations, it is interesting to note that the complete MLE in (4.6) is a special form of the M-estimator in (4.1) where the loss function satisfies

$$\rho(u) = \begin{cases} ||u||^2 & \text{if } \delta_i = 1; \\ 0 & \text{if } \delta_i = 0. \end{cases}$$

Unfortunately this is not a well-defined loss function when the indicator variables δ_i's are unobservable. Fortunately, when $\delta_i = 1$, we might expect that the Y_i is close to the true μ_1. Meanwhile, when $\delta_i = 0$, we might expect that the outlier Y_i is far away from the true μ_1. This viewpoint motivates us to consider a new loss function of the form

$$\rho(u) \approx \begin{cases} u^2 & \text{if } ||u|| \text{ is small for interested data points} \\ \text{bounded} & \text{if } ||u|| \text{ is large for outliers.} \end{cases} \tag{4.7}$$

4.3 The Proposed Robust Estimator

We assume that our data are from the mixture model in (4.4). Our proposed robust estimator of μ_1 is the M-estimator in (4.1), where we propose to consider a new loss function:

$$\rho(u) = \rho_\alpha(u) = \frac{1 - \exp(-\alpha u^2)}{\alpha} \tag{4.8}$$

for some $\alpha > 0$. This new loss function satisfies the properties in (4.7), since $\rho(u) \to u^2$ as $u \to 0$, and $\rho(u) \to 1/\alpha$ as $u \to \infty$. Note that for a given u, the loss function in (4.8) converges to u^2 as $\alpha \to 0$.

For completeness, let us define the loss function

$$\rho_\alpha(u) = \begin{cases} \frac{1-\exp(-\alpha u^2)}{\alpha} & \text{if } \alpha > 0 \\ u^2 & \text{if } \alpha = 0. \end{cases} \tag{4.9}$$

and our proposed M-estimator is given by

$$\hat{\mu}_\alpha = \arg\min_{\mu} \sum_{i=1}^{n} \rho_\alpha(||Y_i - \mu||). \tag{4.10}$$

Before presenting the general theoretical properties of $\hat{\mu}_\alpha$, it is interesting to consider two extreme cases: (i) $\alpha = 0$ and (ii) $\alpha = \infty$.

(i) When $\alpha = 0$, our proposed M-estimator $\hat{\mu}_\alpha$ in (4.18) becomes the classical least square estimator (LSE), which is simply the sample mean in the point estimation context.

(ii) When $\alpha \to \infty$, the loss function $\rho_\alpha(u)$ goes to 0 for any u. In order to be more meaningful, it will be better to ignore the common factor $1/\alpha$ for all data points, and consider the new limiting function

$$\rho^*_{\alpha=\infty}(u) = \lim_{\alpha\to\infty} [\alpha\rho_\alpha(u)] = \begin{cases} 0, & \text{if } u = 0 \\ 1, & \text{if } u \neq 0. \end{cases} \tag{4.11}$$

The corresponding M-estimator under this new loss function can be obtained by fitting lines of observations and choosing the line has the most observations. For one-dimensional data, this is just the mode.

Hence, for one-dimensional data, the M-estimator $\hat{\mu}_\alpha$ changes from the mean to the mode, as α goes from 0 to ∞. Below we will investigate the properties of the M-estimator $\hat{\mu}_\alpha$ for general $0 < \alpha < \infty$ for one-dimensional data. In this case, our proposed M-estimator $\hat{\mu}_\alpha$ in (4.18) becomes

$$\hat{\mu}_N = \arg\min_{\mu} \frac{1}{N} \sum_{i=1}^{N} \rho_\alpha(Y_i - \mu) = \arg\min_{\mu} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\alpha} \left[1 - e^{-\alpha(Y_i-\mu)^2}, \right]. \tag{4.12}$$

which is referred as the L^α estimator below to emphasize that the loss function depends on α.

Let us now investigate the properties of our proposed L^α estimator in the mixed model

in (4.4) for one-dimensional normal distribution. Without loss of generality, we assume the variance $\Sigma_1 = 1$ for the component f_{μ_0}. For the purpose of easy understanding, we will consider two cases, depending on whether there are no corrupted or contaminated data or not.

Theorem 4.1 shows that our proposed L^α estimator $\hat{\mu}_N$ in (4.12) is consistent and asymptotically normal distributed in the model without the corrupted or contaminated data.

Theorem 4.1. *Assume $Y_1, Y_2, \ldots, Y_N$ are iid $N(\mu_0, 1)$. For any given $\alpha > 0$, as $N \to \infty$, we have $\hat{\mu}_N \to \mu_0$ in probability and*

$$\sqrt{N}(\hat{\mu}_N - \mu_0) \to N(0, (2\alpha+1)^3(4\alpha+1)^{-3/2}).$$

Proof: The proof is a straightforward application of the classical technique to prove the asymptotic properties of the MLE for the iid model. Taking derivative of the function $\rho_\alpha(Y-\mu)$ in Eq.(4.18) with respect to μ yields that $\psi(Y,\mu) = 2(Y-\mu)e^{-\alpha(Y-\mu)^2}$. By the law of large numbers, the estimator $\hat{\mu}_N$ is a consistent estimator and by he Central Limit Theorem (CLT), it asymptotically follows a normal distribution with the asymptotic variance $A(\mu_0)^{-2}B(\mu_0)$, where

$$\begin{aligned} A(\mu_0) &= E[-\frac{\partial\psi(Y,\mu_0)}{\partial\mu}] \\ &= -2E\left[-e^{-\alpha(Y-\mu_0)^2} + 2\alpha(Y-\mu_0)^2 e^{-\alpha(Y-\mu_0)^2}\right] \\ &= 2\left[(2\alpha+1)^{-1/2} - 2\alpha/(2\alpha+1)^{3/2}\right] = 2(2\alpha+1)^{-3/2} \end{aligned}$$

and

$$\begin{aligned} B(\mu_0) &= E[\psi(Y,\mu_0)^2] \\ &= 4E[(Y-\mu_0)^2 e^{-2\alpha(Y-\mu_0)^2}] = 4(4\alpha+1)^{-3/2}. \end{aligned}$$

The theorem follows at once by combining the above results together. □

Theorem 4.2 investigate the case when the corrupted or contaminated data exists. It shows that our proposed L^α estimator is not consistent in general, but the bias is bounded above by a constant C that only depends on α and does not depend on the outlier parameters ϵ and Δ.

Theorem 4.2. *Assume $Y_1, Y_2, \ldots, Y_N$ are iid with distribution $(1-\epsilon)N(\mu_0, 1) + \epsilon N(\mu_0 + \Delta, 1)$. For fixed α, Δ, ϵ, we have $\hat{\mu}_N \to \mu^*$ in probability, where*

$$\mu^* = \arg\min_{\mu} E\left[\frac{1 - e^{-\alpha(Y_i-\mu)^2}}{\alpha}\right] \neq \mu_0. \tag{4.13}$$

The bias between loss functions is bounded,

$$|E_{(1-\epsilon)N(\mu_0,1)+\epsilon N(\mu_0+\Delta,1)}(\hat{\mu}_N) - E_{N(\mu_0,1)}(\hat{\mu}_N)| \leq C(\alpha), \tag{4.14}$$

where $C(\alpha)$ only depends on α and does not depend on Δ, ϵ.

Proof: Similar to Theorem 4.1, the sample mean will converge to the population mean, by law of large numbers, we have $\hat{\mu}_N$ converges to μ^*. The bias is given by

$$\begin{aligned}
& E_{(1-\epsilon)N(\mu_0,1)+\epsilon N(\mu_0+\Delta,1)}(\rho_\alpha(Y_i-\mu)) - E_{N(\mu_0,1)}(\rho_\alpha(Y_i-\mu)) \\
= \quad & \frac{1}{\alpha} - \frac{1}{\alpha}\Big[(1-\epsilon)(2\alpha+1)^{-1/2} e^{-\frac{\alpha}{2\alpha+1}(\mu-\mu_0)^2} \\
& + \epsilon(2\alpha+1)^{-1/2} e^{-\frac{\alpha}{2\alpha+1}(\mu-\mu_0-\Delta)^2}\Big] - \left[\frac{1}{\alpha} - \frac{1}{\alpha}(2\alpha+1)^{-1/2}\right] \\
\leq \quad & \frac{1}{\alpha}(2\alpha+1)^{-1/2} \triangleq C(\alpha),
\end{aligned} \tag{4.15}$$

and we also have

$$\begin{aligned}
& E_{(1-\epsilon)N(\mu_0,1)+\epsilon N(\mu_0+\Delta,1)}(\rho_\alpha(Y_i-\mu)) - E_{N(\mu_0,1)}(\rho_\alpha(Y_i-\mu)) \\
& \geq \epsilon\frac{1}{\alpha}(2\alpha+1)^{-1/2} - \epsilon\frac{1}{\alpha}(2\alpha+1)^{-1/2} = 0.
\end{aligned} \tag{4.16}$$

Therefore, Eq.(4.14) holds. □

Theorem 4.3 below extends Theorem 4.2 in the new asymptotic scenario when both $\epsilon = \epsilon_N$ and $\Delta = \Delta_N$ depend on N as $N \to \infty$. For instance, when $\epsilon_N = AN^{-\beta}$ and $\Delta_N = BN^{\gamma}$ for some positive constants A, B, β, and γ, then our theorem shows that our proposed estimator $\hat{\mu}_N$ is asymptotically consistent as long as $\beta > \gamma$.

Theorem 4.3. *Assume* $Y_1, Y_2, \ldots, Y_N$ *are iid with distribution* $(1-\epsilon_N)N(\mu_0, 1)+\epsilon_N N(\mu_0+\Delta_N, 1)$. *If* $\epsilon_N \Delta_N = o(1)$*, then* $\hat{\mu}_N$ *is asymptotically consistent.*

Proof: The proof relies on Theorem 4.2. Since μ^* can not be solved explicitly, Taylor expansion is used to get an approximate solution.

$$E[\rho_\alpha(Y-\mu)] \approx \frac{1}{\alpha} - \frac{1}{\alpha(\alpha+1)^{1/2}} \left\{ (1-\epsilon_N) \left[1 - \frac{\alpha}{2\alpha+1}(\mu-\mu_0)^2 \right] \right.$$
$$\left. +\epsilon_N \left[1 - \frac{\alpha}{2\alpha+1}(\mu-\mu_0-\Delta)^2 \right] \right\}$$

Take derivative with respect to μ, and set to 0, then we have

$$\hat{\mu}_N \approx \mu_0 + \epsilon_N \Delta_N.$$

So if $\epsilon_N \Delta_N = o(1)$, then $\hat{\mu}_N$ is asymptotic consistent. □

4.4 Applications for Regression Models

In this section, we extend the exponential loss function to robustly estimate the parameters in the context of linear regression in the presence of outliers. Assume we observe $(Y_i, X_i) \in \mathbb{R} \times \mathbb{R}^p$ for $i = 1, 2, \ldots, n$, and the generative model for Y_i is given by:

$$Y_i = \begin{cases} X_i^T \beta + \epsilon_i, & \text{with prob. } 1-\pi(x_i) \\ X_i^T \beta + h(X_i) + \tau\epsilon_i, & \text{with prob. } \pi(x_i), \end{cases} \tag{4.17}$$

where $h(X_i) \in \mathbb{R}$ is an unknown function, and ϵ_i are iid with $N(0, \sigma^2)$. Here both standard derivation related parameters, $\sigma > 0$ and $\tau \geq 1$, are unknown. The main objective is to use the observed data $\{(Y_i, X_i)\}_{i=1}^n$ to estimate $\beta \in \mathbb{R}^p$.

When we knew the basis of the nonlinear function h, i.e., $h(X) = \beta_2 h_1(X)$ for the know function $h_1(X)$, then the EM algorithms can be used to derive the MLE estimates of β and β_2. The details of the EM algorithm are discussed in the Appendix. As we will see from our simulation studies, $\hat{\beta}_{LS}$ is not a consistent estimator of β in general.

It is useful to point out the difference between our model in (4.17) with the Huber loss model in (4.3). Let $I_i = 0$ or 1 be the indicator function, which determines where the i-th observation comes from. Our model in (4.17) can be rewritten as

$$Y_i = X_i^T \beta + \Delta_i + \epsilon_i,$$

where

$$\Delta_i = [h(X_i) + (\tau - 1)\epsilon_i] I_i, \text{ for } i = 1, 2, \ldots, n.$$

As compared to the Huber loss model in (4.3), the outlier Δ_i will depend on the X_i and ϵ_i in our context, and thus Huber's M-estimator or Lasso-type argument does not work in our context.

From the concept level, it is straightforward to extend our proposed robust point estimation to robust estimation in the linear regression context. That is, under the model (4.17), our proposed robust estimator of β is the solution of the optimization problem:

$$\hat{\beta}_\alpha = \arg\min_{\mu} \sum_{i=1}^{n} \rho_\alpha(||Y_i - X_i^T \beta||), \tag{4.18}$$

where the loss function $\rho_\alpha(u)$ is given by (4.9).

From the asymptotic theory viewpoint, however, it is difficult to develop a general theoretical theory for our proposed robust estimator $\hat{\beta}_\alpha$ in (4.18) in the linear regression

context. From the practical viewpoint, it is crucial to decide how to choose the best α for a specific problem. One idea is to compute different $\hat{\beta}_\alpha$ with respect to different α's, and then decide the best α for a specific problem under some suitable criterion depending on the inference or prediction context.

Below we will present an efficient numerical algorithm to compute our robust estimator $\hat{\beta}_\alpha$ in (4.18) via the gradient descent algorithm for a given α, and then report our numerical simulation results.

4.4.1 Our Proposed Algorithm for L^α Estimator

We propose to solve the optimization problem (4.18) via the gradient descent algorithm, and it turns out that this can be done by solving a weighted least square regression with a specific weight.

Our proposed algorithm is given as follows when $\alpha > 0$:

Initialize: β and $w = (w_1, \ldots, w_n)$ to be zero vectors, where β is the parameter to be estimated and w is a weight vector with the length n.

While: the iteration loop is less than a pre-set large number K, e.g. $K = 10^4$.

do: (a) Store the current β vector into β_{prev}.

(b) Update the current weight vector w via

$$w_i = \exp(-\alpha(Y_i - X_i^T \beta_{\text{prev}})^2). \tag{4.19}$$

(c) Normalize the weight w by

$$w_i = \frac{w_i}{\sum_{i=1}^n w_i}. \tag{4.20}$$

(d) Fit the weight least square regression between Y_i and X_i with weight w_i.

(e) Compute the estimate β through the weight least square regression.

if: $|\beta - \beta_{\text{prev}}| < \epsilon^*$, e.g., $= 0.0001$, terminate loops and return β.

end: the while loop.

Our numerical experiences suggest that the proposed algorithm seems to converge quickly and work well for our purpose. However, we acknowledge that we do not have a rigorous theoretical proof whether our proposed algorithm always converges to the true solution of the optimization problem (4.18). The challenge is that the loss function $\rho_\alpha(u)$ in (4.9) is not a convex function for $\alpha > 0$, and we are facing a non-convex optimization problem.

4.4.2 Simulation Studies

In this section, we report the simulation properties of our proposed L_α estimator using the numerical algorithm in the previous section. As a comparison, we consider two baseline estimators: the MLE and Huber's estimator. For simplicity, we consider the simple linear regression of Y_i on one-dimensional explanation variable X_i in the presence of outliers.

The details of our simulation setting are as follows:

1. Generate $N = 10^5$ iid x_i's from the interval $[0, 1]$.

2. For each x_i, generate the response y_i's from x_i by the model (4.17) with the following specific functions: $\beta = 1$, and $h(x) = 10x^2$, $\sigma = 5$ and $\pi(x) = \Phi(\alpha_0 + \alpha_1 x)$ with $\alpha_0 = -2.645$ and $\alpha_1 = 1$. Note that this choice of $\pi(x)$ or (α_0, α_1) with the fact of $0 \le x \le 1$ show that $\pi(x) \in [\Phi(-2.645), \Phi(-1.645)] = [0.004, 0.05]$. In other words, the proportion of outliers is in the range of 0.4% and 5%, and the larger x_i's have large chances to yield outliers.

3. For each set of N iid points (x_i, y_i)'s, we can use different estimation methods to obtain a specific numerical estimate $\hat{\beta}$, say, the MLE, Huber's estimation, or use

the numerical algorithm method in the previous section to obtain our proposed L_α estimator.

4. We repeat the above process M times to get M estimates for each estimation method, say, $\hat{\beta}^{(k)}$ for $k = 1, \cdots, M$.

5. Different estimation methods will then be evaluated based on the MSE

$$\frac{1}{M}\sum_{m=1}^{M}(\hat{\beta}^{(m)} - \beta)^2,$$

where the true $\beta = 1$ under our previous setting.

When we estimate β blindly by incorrectly assuming no outliers, the MLE is given by $\hat{\beta} = 1.6076$ (with the standard derivation $sd(\hat{\beta}) = 0.0010$), and the empirical 95% confidence interval of $\hat{\beta}$ is $[1.6077, 1.6077]$. That is, the MLE under the incorrect mode assumption works poorly and is far away from the true parameter value. This is expected, since it mis-specifies the model. As a comparison, Huber's estimator of β is ≈ 1.05 and our proposed L_α estimator is ≈ 1.0001 for certain exponential loss α. In other words, our proposed L_α estimator can be better than the MLE or Huber's estimator in the regression model context.

It is useful to emphasize the importance of the exponential loss parameter α. Consider the case of $\alpha = 0.01$, and we will have $\hat{\beta} = 1.1102$, and when we repeat $K = 100$ times, the standard deviation is $std(\hat{\beta}) = 0.0058$ and the empirical 95% confidence interval is given by $[1.1008, 1.1214]$. As compared to the MLE which can be thought of our proposed L_α estimator with $\alpha = 0$, and the results show that the L_α estimator with $\alpha = 0.01$ is much closer to the true value $\beta = 1$, but the 95% confidence interval still does not cover the true $\beta = 1$ value. This indicates that our proposed L_α estimator with small α has similar properties to the MLE, and can have serious biases in the traditional setting.

Since the estimator $\hat{\beta} = \hat{\beta}_\alpha$ depends on the exponential loss parameter α, we further investigate the relationship between the exponential loss parameter α and the robust estimate

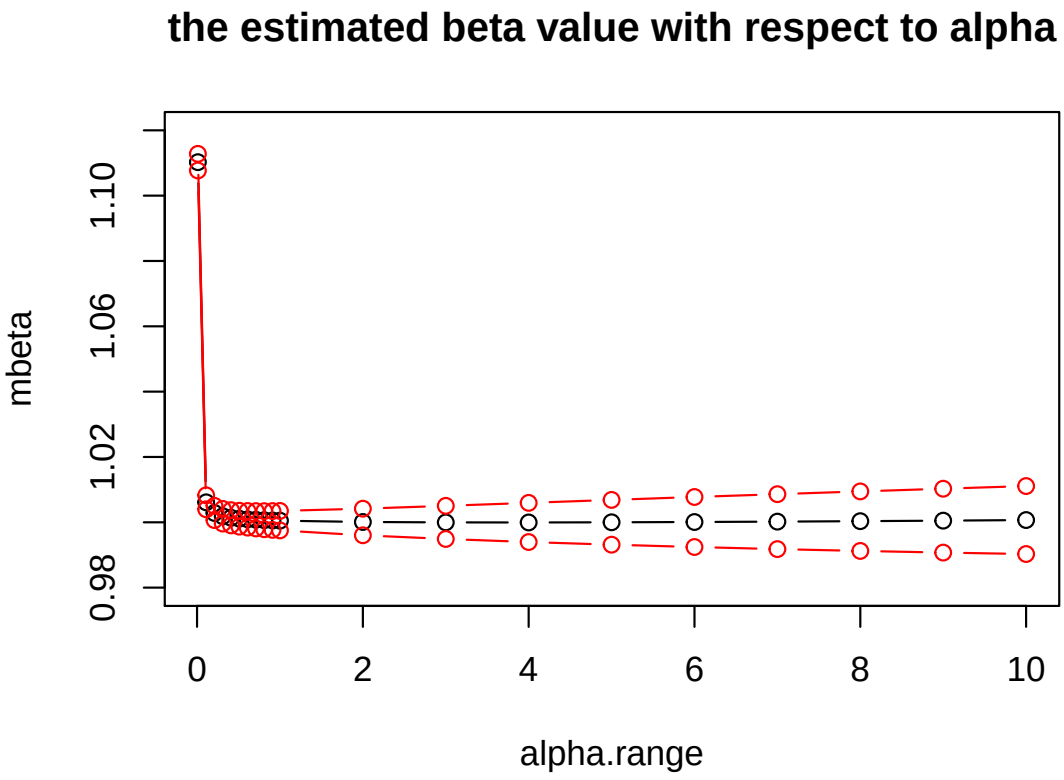

Figure 4.1: The estimated β value with respect to α.

$\hat{\beta}_\alpha$. Figure 4.1 shows how the $\hat{\beta}$ changes with respect to α, where α varies from 0.01 to 10. In addition, Figure 4.2 plots of the MSE of $\hat{\beta}$ as a function of α, which shows that the MSE is minimized at $\alpha = 0.3$.

While we are unable to develop a general guideline to pick up a good exponential loss parameter α, we feel that one might be able to do so by minimizing the MSE if one has some prior knowledge about the outliers. Hopefully such choices of α will have efficient robust properties when the outlier distributions are deviated from the prior knowledge.

4.5 Discussion

In this chapter, we developed a robust estimator in the presence of contaminated data. The properties and the simulation results demonstrate the usefulness of our proposed method

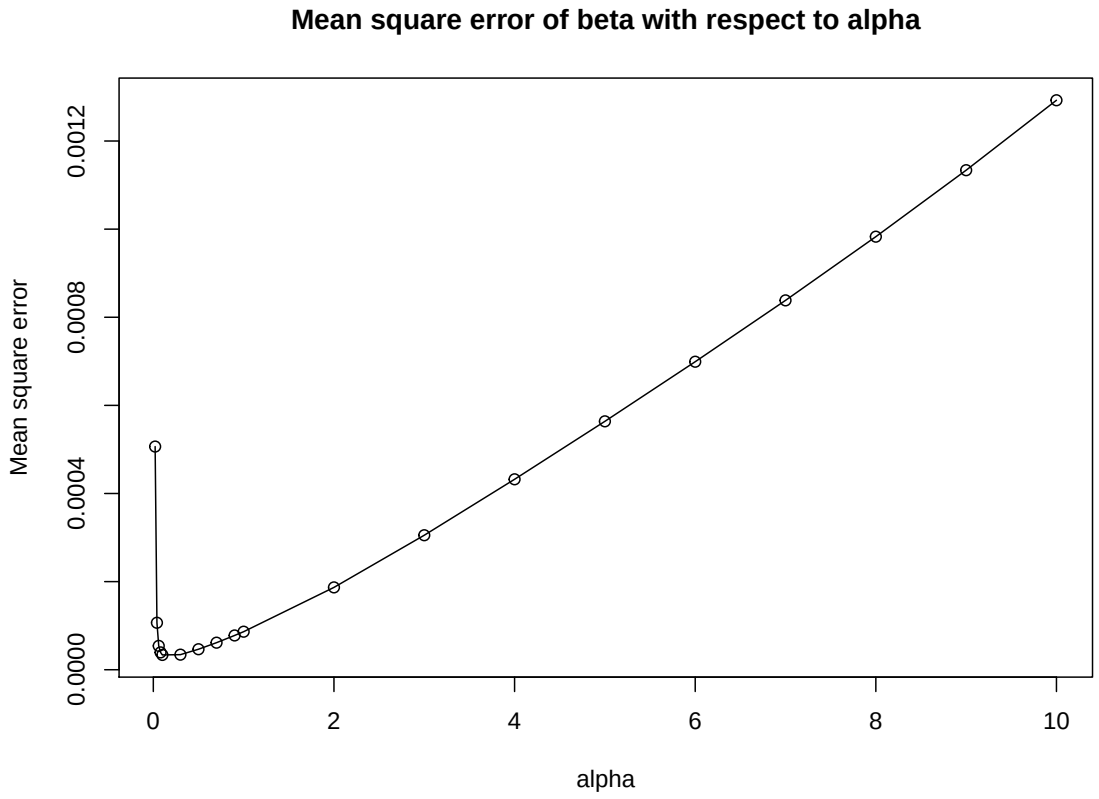

Figure 4.2: The MSE of $\hat{\beta}$ with respect to α.

in point estimation and simple linear regression. These are our preliminary results, as the robust statistical inference is a very interesting topic that deserve further research. One interesting topic is the sequential design of experiment problem: instead of generating iid data points, one is allowed to choose x_i's sequentially by ourselves that allow us to have best possible estimation of β.

Appendix: Derivation of the Linear square estimator

Here we present the EM algorithm when the explanatory variable x_i is one-dimensional. We want to compute the MLE of $(\beta_1, \beta_2, \sigma^2)$. Suppose $\Delta = (\delta_1, \cdots, \delta_N)$ are the latent

variables that determine the component from which the observation originates.

$$Y_{i_1}|(\delta_i = 1) \sim \mathcal{N}(\beta_1 x_i + \beta_2 h(x_i), \sigma^2), \text{ and } Y_{i_2}|(\delta_i = 0) \sim \mathcal{N}(\beta_1 x_i, 1),$$

and

$$Y_i = \delta_i Y_{i_1} + (1 - \delta_i) Y_{i_2},$$

where δ_i follows a binomial distribution $\delta_i \sim Binomial(\Phi(\alpha_0 + \alpha_1 x_i))$. The aim is to estimate the unknown parameters: $(\beta_1, \beta_2, \sigma^2)$. The log-likelihood function of the incomplete data is given by

$$\begin{aligned} l((\beta_1, \beta_2, \sigma^2); \mathbf{y}) &= \sum_{i=1}^{N} \log \left[(\Phi(\alpha_0 + \alpha_1 x_i) \phi \left(\frac{y_i - \beta_1 x_i - \beta_2 h(x_i)}{\sigma} \right) \right. \\ &\quad \left. + (1 - \Phi(\alpha_0 + \alpha_1 x_i)) \phi(y_i - \beta_1 x_i) \right], \end{aligned}$$

and the log-likelihood function of the complete data is given by

$$\begin{aligned} l((\beta_1, \beta_2, \sigma^2); \mathbf{y}, \Delta) &= \sum_{i=1}^{N} \left\{ \log \left[\delta_i \phi \left(\frac{y_i - \beta_1 x_i - \beta_2 h(x_i)}{\sigma} \right) + (1 - \delta_i) \phi(y_i - \beta_1 x_i) \right] \right. \\ &\quad \left. + \delta_i \log(\Phi(\alpha_0 + \alpha_1 x_i)) + (1 - \delta_i) \log(1 - \Phi(\alpha_0 + \alpha_1 x_i)) \right\} \end{aligned}$$

E-step: Given the current estimate $(\beta_1^{(t)}, \beta_2^{(t)}, \sigma^{2(t)})$ of the parameters, the conditional distribution of δ_i is the posterior distribution after taking observations

$$\begin{aligned} \gamma_i^{(t)} &= \mathbf{P}(\delta_i = 1 | \beta_1^{(t)}, \beta_2^{(t)}, \sigma^{2(t)}) \\ &= \frac{\Phi(\alpha_0 + \alpha_1 x_i) \phi \left(\frac{y_i - \beta_1^{(t)} x_i - \beta_2^{(t)} h(x_i)}{\sigma^{(t)}} \right)}{\Phi(\alpha_0 + \alpha_1 x_i) \phi \left(\frac{y_i - \beta_1^{(t)} x_i - \beta_2^{(t)} h(x_i)}{\sigma^{(t)}} \right) + (1 - \Phi(\alpha_0 + \alpha_1 x_i)) \phi(y_i - \beta_1^{(t)} x_i)}. \end{aligned}$$

The Q-function in the expectation step:

$$Q(\beta_1,\beta_2,\sigma^2|\beta_1^{(t)},\beta_2^{(t)},\sigma^{2(t)}) = \mathbf{E}_{\Delta|\mathbf{y},\beta_1^{(t)},\beta_2^{(t)},\sigma^{2(t)}}[l((\beta_1,\beta_2,\sigma^2);\mathbf{y},\Delta)]$$

$$= \sum_{i=1}^{N} \mathbf{E}_{\Delta|\mathbf{y},\beta_1^{(t)},\beta_2^{(t)},\sigma^{2(t)}}[l((\beta_1,\beta_2,\sigma^2);y_i,\delta_i)]$$

$$= \sum_{i=1}^{N} \left[\gamma_i^{(t)} l((\beta_1,\beta_2,\sigma^2);y_i,\delta_i=1) + (1-\gamma_i^{(t)}) l((\beta_1,\beta_2,\sigma^2);y_i,\delta_i=0)\right]$$

$$= \sum_{i=1}^{N} \left\{\gamma_i^{(t)} \left[\log\left(\phi\left(\frac{y_i-\beta_1 x_i-\beta_2 h(x_i)}{\sigma}\right)\right) + \log(\Phi(\alpha_0+\alpha_1 x_i))\right]\right.$$

$$\left. +(1-\gamma_i^{(t)}) \left[\log\left(\phi(y_i-\beta_1 x_i)\right) + \log(1-\Phi(\alpha_0+\alpha_1 x_i))\right]\right\}$$

$$= \sum_{i=1}^{N} \left\{\gamma_i^{(t)} \left[-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y_i-\beta_1 x_i-\beta_2 h(x_i))^2 + \log(\Phi(\alpha_0+\alpha_1 x_i))\right]\right.$$

$$\left. +(1-\gamma_i^{(t)}) \left[-\frac{1}{2}\log(2\pi) - \frac{1}{2}(y_i-\beta_1 x_i)^2 + \log(1-\Phi(\alpha_0+\alpha_1 x_i))\right]\right\}.$$

M-step:

$$(\beta_1^{(t+1)},\beta_2^{(t+1)},\sigma^{2(t+1)}) = \arg\max_{\beta_1,\beta_2,\sigma^2} Q(\beta_1,\beta_2,\sigma^2|\beta_1^{(t)},\beta_2^{(t)},\sigma^{2(t)})$$

$$= \sum_{i=1}^{N} \left\{\gamma_i^{(t)} \left[-\frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y_i-\beta_1 x_i-\beta_2 h(x_i))^2\right]\right.$$

$$\left. +(1-\gamma_i^{(t)}) \left[-\frac{1}{2}(y_i-\beta_1 x_i)^2\right]\right\}.$$

We then use gradient descent to find the local maximum of the Q function.

$$\begin{pmatrix} {\beta_1^{(t)}}^{(k+1)} \\ {\beta_2^{(t)}}^{(k+1)} \\ {\sigma^2}^{(t)(k+1)} \end{pmatrix} = \begin{pmatrix} {\beta_1^{(t)}}^{(k)} \\ {\beta_2^{(t)}}^{(k)} \\ {\sigma^2}^{(t)(k)} \end{pmatrix} + \zeta \begin{pmatrix} \sum_{i=1}^{N} \left[\gamma_i^{(t)} \frac{x_i}{{\sigma^2}^{(t)(k)}} (y_i - {\beta_1^{(t)}}^{(k)} x_i - {\beta_2^{(t)}}^{(k)} h(x_i)) + (1 - \gamma_i^{(t)}) x_i (y_i - {\beta_1^{(t)}}^{(k)} x_i) \right] \\ \sum_{i=1}^{N} \gamma_i^{(t)} \frac{h(x_i)}{{\sigma^2}^{(t)(k)}} (y_i - {\beta_1^{(t)}}^{(k)} x_i - {\beta_2^{(t)}}^{(k)} h(x_i)) \\ \sum_{i=1}^{N} \gamma_i^{(t)} \left[-\frac{1}{2{\sigma^2}^{(t)(k)}} + \frac{1}{2({\sigma^2})^{2(t)(k)}} (y_i - {\beta_1^{(t)}}^{(k)} x_i - {\beta_2^{(t)}}^{(k)} h(x_i))^2 \right] \end{pmatrix},$$

where ζ is the step-size. After picking some initial guess of the three parameters, we iterate until it converges. The converged value is set to be $(\beta_1^{(t+1)}, \beta_2^{(t+1)}, {\sigma^2}^{(t+1)})$.

CHAPTER 5
CONCLUSIONS AND FUTURE RESEARCH

This thesis investigates effective data integration techniques in four different contexts: (i) online monitoring of large-scale data streams, (ii) consensus sequential detection over distributed networks, (iii) combining different patients' responses to assess the treatment effects of new drugs, and (iv) robust statistical inference in the presence of contaminated data.

Our research opens several further research directions on the development of effective data integration techniques in more complicated data sources or models.

- **Distributed spatio-temporal data**. The first two chapters of this thesis focus on the scenario of distributed sensor networks when the data are independent and identically distributed (conditional on a given hypothesis). It will be interesting to consider more complicated spatio-temporal models and develop efficient data integration techniques that can take into account of spatial or temporal correlations. For instance, it will be interesting to develop scalable quickest change detection schemes when the change occurs to the means of some local sensors or the correlation structures of some subsets of local sensors, or to develop efficient sequential tests in the distributed network system when each local sensor can only communicate with its immediate neighborhood sensors.

- **Modern asymptotic theory for sequential decisions.** In the classical asymptotic regime for sequential decision problems, one fixes the dimension of the data, as the expected sample size goes to ∞. In Chapters 1, 2 and 4 of this thesis, our results deal with the modern asymptotic regime, e.g., where both the dimension of data and the expected sample size go to ∞ simultaneously in a suitable rate, or when both

outlier proportion and outlier size vary simultaneously as the sample size goes to ∞. We hope more theoretical results for sequential decision problems can be developed under the modern asymptotic regime. It is very challenging to develop asymptotic theories for sequential decisions under the high dimensional setting, and based on our research experience, one possible approach is first to develop some rough non-asymptotic bounds that depends on both dimension and expected sample size, and then to investigate their behaviors of these non-asymptotic bounds under the modern asymptotic regime.

- **Statistical efficiency versus computational efficiency.** Statistical efficiency has been playing a central role in the subfield of sequential analysis and change-point detection since Wald's famous SPRT paper in 1945, and in order to develop optimality or asymptotic optimality theories, much research in the subfield often needs to make strong assumptions of i.i.d. data. In many real-world applications of large-scale data, however, one might be willing to sacrifice a little statistical efficiency to achieve much improved computational efficiency, e.g., distributed procedures/systems might be favored over centralized procedures/systems. If such a case, it will be interesting to investigate how to better balance the tradeoff between statistical efficiency and computational efficiency. We feel that data integration techniques will become crucial tools, and hopefully this thesis can motivate further research along this direction.

REFERENCES

[1] M. S. Albert, S. T. DeKosky, D. Dickson, B. Dubois, H. H. Feldman, N. C. Fox, A. Gamst, D. M. Holtzman, W. J. Jagust, R. C. Petersen, *et al.*, "The diagnosis of mild cognitive impairment due to alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease," *Alzheimer's & dementia: The journal of the alzheimer's association*, vol. 7, no. 3, pp. 270–279, 2011.

[2] S. Appadwedula, V. V. Veeravalli, and D. L. Jones, "Energy-efficient detection in sensor networks," *IEEE j. sel. areas commun.*, vol. 23, pp. 693–702, 2005.

[3] T. Banerjee and V. V. Veeravalli, "Data-efficient quickest change detection in sensor networks," *IEEE trans. signal processing*, vol. 63, pp. 3727–3735, 2015.

[4] M. Basseville and I. V. Nikiforov, *Detection of abrupt changes: Theory and applications*. Englewood Cliffs, Prentice-Hall, 1993.

[5] A. Basu and A. Manca, "Regression estimators for generic health-related quality of life and quality-adjusted life years," *Medical decision making*, vol. 32, no. 1, pp. 56–69, 2012.

[6] R. S. Blum, S. A. Kassam, and H. V. Poor, "Distributed detection with multiple sensors i. advanced topics," *In proc. of the IEEE*, vol. 85, no. 1, pp. 64–79, 1997.

[7] E. J. Candès, "Modern statistical estimation via oracle inequalities," *Acta numerica*, vol. 15, pp. 257–325, 2006.

[8] J. M. Cedarbaum, M. Jaros, C. Hernandez, N. Coley, S. Andrieu, M. Grundman, and B. Vellas, "Rationale for use of the clinical dementia rating sum of boxes as a primary outcome measure for alzheimer's disease clinical trials," *Alzheimer's & dementia: The journal of the alzheimer's association*, vol. 9, no. 1, S45–S55, 2013.

[9] Y. Chen, C. Caramanis, and S. Mannor, "Robust sparse regression under adversarial corruption," in *International conference on machine learning*, 2013, pp. 774–782.

[10] F. R. Chung, *Spectral graph theory*, 92. American Mathematical Soc., 1997.

[11] N. Coley, S. Andrieu, M. Jaros, M. Weiner, J. Cedarbaum, and B. Vellas, "Suitability of the clinical dementia rating-sum of boxes as a single primary endpoint for alzheimer's disease trials," *Alzheimer's & dementia: The journal of the alzheimer's association*, vol. 7, no. 6, pp. 602–610, 2011.

[12] F. Cribari-Neto and A. Zeileis, "Beta regression in R," *Journal of statistical software*, vol. 34, pp. 1–24, 2010.

[13] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.

[14] J. Fan and S. K. Lin, "Test of significance when data are curves," *Journal of american statistical association*, vol. 93, pp. 1007–1021, 1998.

[15] G. Fellouris, "Asymptotically optimal parameter estimation under communication constraints," *Ann. statist.*, vol. 40, no. 4, pp. 2239–2265, 2012.

[16] D. Ferrari and Y. Yang, "Maximum lq-likelihood estimation," *The annals of statistics*, vol. 38, no. 2, pp. 753–783, 2010.

[17] S. Ferrari and F. Cribari-Neto, "Beta regression for modelling rates and proportions," *Journal of applied statistics*, vol. 31, no. 7, pp. 799–815, 2004.

[18] G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs, *Longitudinal data analysis*. CRC Press, 2008.

[19] G. M. Fitzmaurice, N. M. Laird, and J. H. Ware, *Applied longitudinal analysis*. John Wiley & Sons, 2012, vol. 998.

[20] B. E. Flinchbaugh and B Chandrasekaran, "A theory of spatio-temporal aggregation for vision," *Artificial intelligence*, vol. 17, no. 1-3, pp. 387–407, 1981.

[21] C. D. Fuh and Y. Mei, "Quickest change detection and kullback-leibler divergence for two-state hidden Markov models," *IEEE trans. signal processing*, vol. 63, pp. 4866–4878, 2015.

[22] A. Galstyan, B. Krishnamachari, K. Lerman, and S. Pattem, "Distributed online localization in sensor networks using a moving target," in *3rd international symposium on inf. proc. in sensor networks (IPSN)*, 2004, pp. 61–70.

[23] A. Gelman, D. Lee, and J. Guo, "Stan: A probabilistic programming language for bayesian inference and optimization," *Journal of educational and behavioral statistics*, vol. 40, no. 5, pp. 530–543, 2015.

[24] A Giménez-Arnau, I Izquierdo, and M Maurer, "The use of a responder analysis to identify clinically meaningful differences in chronic urticaria patients following placebo-controlled treatment with rupatadine 10 and 20 mg," *Journal of the european academy of dermatology and venereology*, vol. 23, no. 9, pp. 1088–1091, 2009.

[25] K. Gimpel, D. Das, and N. A. Smith, "Distributed asynchronous online learning for natural language processing," in *Proc. of the 14th conference on computational natural language learning*, Association for Computational Linguistics, 2010, pp. 213–222.

[26] J. Glaz, J. Naus, and S. Wallenstein, *Scan statistics*. Springer-Verlag, New York, 2001.

[27] L. Gordon and M. Pollak, "An efficient sequential nonparametric scheme for detecting a change of distribution," *Ann. statist.*, vol. 22, pp. 763–804, 1994.

[28] H. R. Hashemi and I. B. Rhodes, "Decentralized sequential detection," *IEEE trans. on inf. theory*, vol. 35, no. 3, pp. 509–520, 1989.

[29] D. Hedeker and R. D. Gibbons, *Longitudinal data analysis*. John Wiley & Sons, 2006, vol. 451.

[30] S.-F. Huang, C.-K. Liu, C.-C. Chang, and C.-Y. Su, "Sensitivity and specificity of executive function tests for alzheimer's disease," *Applied neuropsychology: Adult*, vol. 24, no. 6, pp. 493–504, 2017.

[31] P. J. Huber, "The behavior of maximum likelihood estimates under nonstandard conditions," in *Proceedings of the fifth berkeley symposium on mathematical statistics and probability*, Berkeley, CA, vol. 1, 1967, pp. 221–233.

[32] ——, "The 1972 Wald lecture robust statistics: A review," *The annals of mathematical statistics*, pp. 1041–1067, 1972.

[33] ——, "Robust statistics," in *International encyclopedia of statistical science*, Springer, 2011, pp. 1248–1251.

[34] M. Hunger, A. Döring, and R. Holle, "Longitudinal beta regression models for analyzing health-related quality of life scores over time," *BMC medical research methodology*, vol. 12, no. 1, p. 144, 2012.

[35] K. Ito and M. M. Hutmacher, "Predicting the time to clinically worsening in mild cognitive impairment patients and its utility in clinical trial design by modeling a longitudinal clinical dementia rating sum of boxes from the adni database," *Journal of alzheimer's disease*, vol. 40, no. 4, pp. 967–979, 2014.

[36] D. Jakovetic, J. M. F. Moura, and J. Xavier, "Distributed detection over noisy networks: Large deviations analysis," *IEEE trans. on signal process.*, vol. 60, no. 8, pp. 4306–4320, 2012.

[37] S. Kar, S. Aldosari, and J. M. F. Moura, "Topology for distributed inference on graphs," *IEEE trans. on signal process.*, vol. 56, no. 6, pp. 2609–2613, 2008.

[38] S. Kar and J. M. F. Moura, "Sensor networks with random links: Topology design for distributed consensus," *IEEE trans. on signal process.*, vol. 56, no. 7, pp. 3315–3326, 2008.

[39] S. Kar and J. M. Moura, "Consensus based detection in sensor networks: Topology optimization under practical constraints," in *Proc. 1st intl. wrkshp. inform. theory sensor networks*, 2007.

[40] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art," *Information fusion*, vol. 14, no. 1, pp. 28–44, 2013.

[41] M. Kulldorff, "Prospective time-periodic geographic disease surveillance using a scan statistic," *J. r. stat. soc. ser. a*, vol. 164, pp. 61–72, 2001.

[42] T. L. Lai, "Sequential change-point detection in quality control and dynamical systems (with discussion)," *J. r. stat. soc. ser. b stat. methodol.*, vol. 57, pp. 613–658, 1995.

[43] ——, "Sequential analysis: Some classical problems and new challenges," *Statist. sinica*, vol. 11, pp. 303–408, 2001.

[44] N. M. Laird and J. H. Ware, "Random-effects models for longitudinal data," *Biometrics*, pp. 963–974, 1982.

[45] C. Lévy-Leduc and F. Roueff, "Detection and localization of change-points in high-dimensional network traffic data," *Ann. appl. stat.*, vol. 3, pp. 637–662, 2009.

[46] S. Li and X. Wang, "Order-2 asymptotic optimality of the fully distributed sequential hypothesis test," *Arxiv preprint arxiv:1606.04203*, 2016.

[47] K.-Y. Liang and S. L. Zeger, "Longitudinal data analysis using generalized linear models," *Biometrika*, vol. 73, no. 1, pp. 13–22, 1986.

[48] Y. Liang, L. Lai, and J. Halloran, "Distributed algorithm for collaborative detection in cognitive radio networks," in *2009 47th annual allerton conference on communication, control, and computing (allerton)*, 2009, pp. 394–399.

[49] ——, "Distributed cognitive radio network management via algorithms in probabilistic graphical models," *IEEE journal on selected areas in commun.*, vol. 29, no. 2, pp. 338–348, 2011.

[50] K. Liu, Y. Mei, and J. Shi, "An adaptive sampling strategy for online high-dimensional process monitoring," *Technometrics*, vol. 57, pp. 305–319, 2015.

[51] P.-L. Loh, "Statistical consistency and asymptotic normality for high-dimensional robust m-estimators," *The annals of statistics*, vol. 45, no. 2, pp. 866–896, 2017.

[52] G. Lorden, "Procedures for reacting to a change in distribution," *Ann. math. statist.*, vol. 42, pp. 1897–1908, 1971.

[53] G. Lorden and M. Pollak, "Sequential change-point detection procedures that are nearly optimal and computationally simple," *Sequential analysis*, vol. 27, pp. 476–512, 2008.

[54] P. D. Lorenzo and S. Barbarossa, "Distributed estimation and control of algebraic connectivity over random graphs," *IEEE trans. on signal process.*, vol. 62, no. 21, pp. 5615–5628, 2014.

[55] V. Matta, P. Braca, S. Marano, and A. H. Sayed, "Diffusion-based adaptive distributed detection: Steady-state performance in the slow adaptation regime," *IEEE trans. on inf. theory*, vol. 62, no. 8, pp. 4710–4732, 2016.

[56] G. M. McKhann, D. S. Knopman, H. Chertkow, B. T. Hyman, C. R. Jack, C. H. Kawas, W. E. Klunk, W. J. Koroshetz, J. J. Manly, R. Mayeux, *et al.*, "The diagnosis of dementia due to alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease," *Alzheimer's & dementia: The journal of the alzheimer's association*, vol. 7, no. 3, pp. 263–269, 2011.

[57] Y. Mei, "Information bounds and quickest change detection in decentralized decision systems," *IEEE trans. inform. theory*, vol. 51, pp. 2669–2681, 2005.

[58] ——, "Asymptotic optimality theory for decentralized sequential hypothesis testing in sensor networks," *IEEE trans. on inf. theory*, vol. 54, no. 5, pp. 2072–2089, 2008.

[59] ——, "Efficient scalable schemes for monitoring a large number of data streams," *Biometrika*, vol. 97, pp. 419–433, 2010.

[60] ——, "Quickest detection in censoring sensor networks," in *Proceedings of IEEE international symposium on information theory (ISIT)*, 2011, pp. 2148–2152.

[61] M. Mesbahi and M. Egerstedt, *Graph theoretic methods in multiagent networks*. Princeton University Press, 2010.

[62] R. A. Moore, O. A. Moore, S. Derry, P. M. Peloso, A. R. Gammaitoni, and H. Wang, "Responder analysis for pain relief and numbers needed to treat in a meta-analysis of etoricoxib osteoarthritis trials: Bridging a gap between clinical trials and clinical practice," *Annals of the rheumatic diseases*, vol. 69, no. 2, pp. 374–379, 2010.

[63] R. Moore, N Cai, V Skljarevski, and T. Tölle, "Duloxetine use in chronic painful conditions–individual patient data responder analysis," *European journal of pain*, vol. 18, no. 1, pp. 67–75, 2014.

[64] G. V. Moustakides, "Optimal stopping times for detecting changes in distributions," *Ann. statist.*, vol. 14, pp. 1379–1387, 1986.

[65] J. Neyman, "Smooth test for goodness-of-fit," *Skand. aktuarietidskr.*, vol. 20, pp. 149–199, 1937.

[66] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, pp. 100–115, 1954.

[67] M. Pollak, "Optimal detection of a change in distribution," *Ann. statist.*, vol. 13, pp. 206–227, 1985.

[68] ——, "Average run lengths of an optimal method of detecting a change in distribution," *Ann. statist.*, vol. 15, pp. 749–779, 1987.

[69] H. V. Poor and O. Hadjiliadis, *Quickest detection*. Cambridge Univ. Press, New York, 2009.

[70] Y. Qin and C. E. Priebe, "Maximum lq-likelihood estimation via the expectation-maximization algorithm: A robust estimation of mixture models," *Journal of the american statistical association*, vol. 108, no. 503, pp. 914–928, 2013.

[71] M. Rabbat and R. Nowak, "Distributed optimization in sensor networks," in *3rd international symposium on inf. proc. in sensor networks (IPSN)*, 2004, pp. 20–27.

[72] M. G. Rabbat and R. D. Nowak, "Decentralized source localization and tracking [wireless sensor networks]," in *2004 IEEE international conference on acoustics, speech, and signal process.*, vol. 3, 2004, iii–921–4 vol.3.

[73] C. Rago, P. Willett, and Y. Bar-Shalom, "Censoring sensors: A low-communication-rate scheme for distributed detection," *IEEE trans. aerosp. electon. syst.*, vol. 32, pp. 554–568, 1996.

[74] S. W. Roberts, "A comparison of some control chart procedures," *Technometrics*, vol. 8, pp. 411–430, 1966.

[75] A. K. Sahu and S. Kar, "Distributed sequential detection for gaussian shift-in-mean hypothesis testing," *IEEE trans.s on signal process.*, vol. 64, no. 1, pp. 89–103, 2016.

[76] A. K. Sahu and S. Kar, "Recursive distributed detection for composite hypothesis testing: Nonlinear observation models in additive gaussian noise," *IEEE trans. on inf. theory*, 2017.

[77] M. N. Samtani, N. Raghavan, G. Novak, P. Nandy, and V. A. Narayan, "Disease progression model for clinical dementia rating–sum of boxes in mild cognitive impairment and alzheimer's subjects from the alzheimer's disease neuroimaging initiative," *Neuropsychiatric disease and treatment*, vol. 10, p. 929, 2014.

[78] W. A. Shewhart, *Economic control of quality of manufactured product*. D Van Norstrand, New York., 1931.

[79] A. N. Shiryaev, "On optimum methods in quickest detection problems," *Theory probab. appl.*, vol. 8, pp. 22–46, 1963.

[80] D. Siegmund, *Sequential analysis: Tests and confidence intervals*. Springer, New York, 1985.

[81] S. M. Snapinn and Q. Jiang, "Responder analyses and the assessment of a clinically relevant treatment effect," *Trials*, vol. 8, no. 1, p. 31, 2007.

[82] T. Sorensen and S. Vasishth, "Bayesian linear mixed models using stan: A tutorial for psychologists, linguists, and cognitive scientists," *Arxiv preprint arxiv:1506.06201*, 2015.

[83] A. Tartakovsky, I. Nikiforov, and M. Basseville, *Sequential analysis: Hypothesis testing and changepoint detection*. Boca Raton : Taylor & Francis, 2015.

[84] A. G. Tartakovsky, B. L. Rozovskiia, R. B. Blazeka, and H. Kim, "Detection of intrusions in information systems by sequential change-point methods (with discussions)," *Statistical methodology*, vol. 3, pp. 252–340, 2006.

[85] W. P. Tay, J. N. Tsitsiklis, and M. Z. Win, "Asymptotic performance of a censoring sensor network," *IEEE trans. inform. theory*, vol. 53, pp. 4191–4209, 2007.

[86] C. Tekin, S. Zhang, and M. van der Schaar, "Distributed online learning in social recommender systems," *IEEE journal of selected topics in signal process.*, vol. 8, no. 4, pp. 638–652, 2014.

[87] T. Uryniak, I. S. Chan, V. V. Fedorov, Q. Jiang, L. Oppenheimer, S. M. Snapinn, C.-H. Teng, and J. Zhang, "Responder analyses: A PhRMA position paper," *Statistics in biopharmaceutical research*, vol. 3, no. 3, pp. 476–487, 2011.

[88] V. V. Veeravalli, T. Basar, and H. V. Poor, "Decentralized sequential detection with a fusion center performing the sequential test," in *1992 american control conference*, 1992, pp. 1177–1181.

[89] J. Verkuilen and M. Smithson, "Mixed and mixture regression models for continuous bounded responses using the beta distribution," *Journal of educational and behavioral statistics*, vol. 37, no. 1, pp. 82–113, 2012.

[90] R. Viswanathan and P. K. Varshney, "Distributed detection with multiple sensors i. fundamentals," *In proc. of the IEEE*, vol. 85, no. 1, pp. 54–63, 1997.

[91] A. Wald, "Sequential tests of statistical hypotheses," *The annals of mathematical statistics*, vol. 16, no. 2, pp. 117–186, 1945.

[92] Y. Wang and Y. Mei, "Large-scale multi-stream quickest change detection via shrinkage post-change estimation," *IEEE trans. inform. theory*, vol. 61, pp. 6926–6938, 2015.

[93] M. W. Weiner, D. P. Veitch, P. S. Aisen, L. A. Beckett, N. J. Cairns, R. C. Green, D. Harvey, C. R. Jack, W. Jagust, E. Liu, *et al.*, "The alzheimer's disease neuroimaging initiative: A review of papers published since its inception," *Alzheimer's & dementia: The journal of the alzheimer's association*, vol. 9, no. 5, e111–e194, 2013.

[94] M. M. Williams, M. Storandt, C. M. Roe, and J. C. Morris, "Progression of alzheimer's disease as measured by clinical dementia rating sum of boxes scores," *Alzheimer's & dementia: The journal of the alzheimer's association*, vol. 9, no. 1, S39–S44, 2013.

[95] L. Xiao and S. Boyd, "Fast linear iterations for distributed averaging," *Systems & control letters*, vol. 53, no. 1, pp. 65–78, 2004.

[96] Y. Xie, J. Huang, and R. Willett, "Changepoint detection for high-dimensional time series with missing data," *IEEE journal of selected topics in signal processing*, vol. 7, pp. 12–27, 2013.

[97] Y. Xie and D. Siegmund, "Sequential multi-sensor change-point detection," *Ann. stat.*, vol. 41, pp. 670–692, 2013.

[98] S. X. Xu, M. N. Samtani, A. Russu, O. J. Adedokun, M. Lu, K. Ito, B. Corrigan, S. Raje, H. R. Brashear, S. Styren, *et al.*, "Alzheimer's disease progression

model using disability assessment for dementia scores from bapineuzumab trials," *Alzheimer's & dementia: Translational research & clinical interventions*, vol. 1, no. 2, pp. 141–149, 2015.

[99] F. Yan, S. Sundaram, S. V. N. Vishwanathan, and Y. Qi, "Distributed autonomous online learning: Regrets and intrinsic privacy-preserving properties," *IEEE trans. on knowl. and data eng.*, vol. 25, no. 11, pp. 2483–2493, 2013.

[100] Y. Zhang, D. Sow, D. Turaga, and M. van der Schaar, "A fast online learning algorithm for distributed mining of bigdata," *Acm sigmetrics performance evaluation review*, vol. 41, no. 4, pp. 90–93, 2014.

[101] Q. Zhou, D. Li, S. Kar, L. Huie, H. V. Poor, and S. Cui, "Learning-based distributed detection-estimation in sensor networks with unknown sensor defects," *Arxiv preprint arxiv:1510.02371*, 2015.

[102] D. Zimprich, "Modeling change in skewed variables using mixed beta regression models," *Research in human development*, vol. 7, no. 1, pp. 9–26, 2010.